Unsupervised Ensemble Learning with Dependent Classifiers

Ariel Jaffe, Ethan Fetaya, Boaz Nadler
Weizmann Institute of Science

Tingting Jiang, Yuval Kluger
Yale University School of Medicine

Abstract

In unsupervised ensemble learning, one obtains predictions from multiple sources or classifiers, yet without knowing the reliability and expertise of each source, and with no labeled data to assess it. The task is to combine these possibly conflicting predictions into an accurate meta-learner. Most works to date assumed perfect diversity between the different sources, a property known as conditional independence. In realistic scenarios, however, this assumption is often violated, and ensemble learners based on it can be severely sub-optimal. The key challenges we address in this paper are: (i) how to detect, in an unsupervised manner, strong violations of conditional independence; and (ii) construct a suitable meta-learner. To this end we introduce a statistical model that allows for dependencies between classifiers. Based on this model, we develop novel unsupervised methods to detect strongly dependent classifiers, better estimate their accuracies, and construct an improved meta-learner. Using both artificial and real datasets, we showcase the importance of taking classifier dependencies into account and the competitive performance of our approach.

1 Introduction

In recent years unsupervised ensemble learning has become increasingly popular. In multiple application domains, one obtains the predictions, over a large set of unlabeled instances, of an ensemble of different experts or classifiers with unknown reliability. Common tasks are combining these possibly conflicting predictions into an accurate meta-learner, and assessing the accuracy of the experts, both without any labeled data.

A leading example is crowdsourcing, whereby a tedious labeling task is distributed to many annotators. Computational biology is another field where unsupervised ensemble learning is of increasing interest. Here, several recent works [1, 3, 8, 17] propose to solve difficult biological prediction tasks by applying multiple algorithms and merging their results. Additional examples of unsupervised ensemble learning appear, among others, in medicine [15] and decision science [20].

Perhaps the first to address unsupervised ensemble learning were Dawid and Skene [6]. A key assumption in their work was of perfect diversity between the different classifiers. Namely, their labeling errors were assumed statistically independent of each other. This property, known as *conditional independence* is illustrated in the graphical model of Fig. 1 (left). In [6], Dawid and Skene proposed to estimate the model parameters, i.e. the accuracies of the different classifiers, by the EM procedure on the non-convex likelihood function. With the increasing popularity of crowdsourcing and other unsupervised ensemble learning applications, there has been a surge of interest in this line of work, and multiple extensions of it [14,21,23,25,26]. As the quality of the solution found by the EM algorithm critically depends on its starting point, several recent works derived computationally efficient spectral methods to suggest a good initial guess [2, 10, 11, 18]. There have also been works employing a Bayesian approach, in particular for unsupervised ordinal prediction problems, see e.g. [12, 13].

Despite its popularity and usefulness, the model of Dawid and Skene has several limitations. One notable limitation is its assumption that all instances are equally difficult, with each classifier having the same probability of error over all instances. This issue was addressed among others, by Whitehill et. al. [26] who introduced a model of instance difficulty, and by Tian et. al. [24] who proposed a model where instances are divided into groups, and the expertise of each classifier is group dependent.

Appearing in Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS) 2016, Cadiz, Spain. JMLR: W&CP volume 51. Copyright 2016 by the authors.

A second limitation, at the focus of our work, is the assumption of perfect conditional independence between all classifiers. As we illustrate below, this assumption may be strongly violated in real-world scenarios. Furthermore, as shown in Sec. 5, neglecting classifier dependencies may yield quite sub-optimal predictions. Yet, to the best of our knowledge, relatively few works have attempted to address this important issue.

To handle classifier dependencies, Donmez et. al. [7] proposed a model with pairwise interactions between all classifier outputs. However, they noted that empirically, their model did not yield more accurate predictions. Platanios et. al. [19] developed a method to estimate the error rates of possibly dependent classifiers. Their method is based on analyzing the agreement rates between groups of classifiers, together with a prior on weak dependence amongst them.

The present work is partly motivated by the ongoing somatic mutation DREAM (Dialogue for Reverse Engineering Assessments and Methods) challenge, a sequence of open competitions for detecting irregularities in the DNA string. This is a real-world example of unsupervised ensemble learning, where participants in this competition are given access to the predictions of more than 100 different classifiers, over more than 100,000 instances. These classifiers were constructed by various labs worldwide, each employing their own biological knowledge and possibly proprietary labeled data. The task is to construct, in an unsupervised fashion, an accurate ensemble learner.

In figure 2 we present the empirical *conditional* covariance matrix between different classifiers in one of the databases of the DREAM challenge, for which ground truth labels have been disclosed. Under the conditional independence assumption, the population conditional covariance between every two classifiers should be exactly zero. Figure 2, in contrast, exhibits strong dependencies between groups of classifiers. In this paper we show the importance of taking these dependencies into account.

Unsupervised ensemble learning in the presence of possibly dependent classifiers raises the following two key challenges: (i) detect, in an unsupervised manner, strong violations of conditional independence; and (ii) construct a suitable meta-learner.

To cope with these challenges, in Sec. 2 we introduce a new model for the joint distribution of all classifiers which allows for dependencies between them through an intermediate layer of latent variables. This generalizes the model of Dawid and Skene, and allows for groups of strongly correlated classifiers, as observed for example in the DREAM data.

In Sec. 3 we devise a simple algorithm to detect subsets of strongly dependent classifiers using only their predictions and no labeled data. This is done by exploiting the structural low-rank properties of the classifiers' covariance matrix. Figure 3 shows our resulting estimate for deviations from conditional independence on the same data as figure 2. Comparing the two figures illustrates the ability of our method to detect strong dependencies with no labeled data.

In Sec. 4 we propose methods to better estimate the accuracies of the classifiers and construct an improved meta-learner, both in the presence of strong dependencies between some of the classifiers. Finally, in Sec. 5 we illustrate the competitive performance of our modified ensemble-learner on artificial data, several UCI datasets and three datasets from the DREAM challenge. These empirical results showcase the limitations of the strict conditional independence model, and highlight the importance of modeling the statistical dependencies between different classifiers in unsupervised ensemble learning scenarios.

2 Problem Setup

Notations. Consider a binary classification problem with instance space $\mathcal{X}$ and output space $\mathcal{Y} = \{-1, 1\}$. A labeled instance $(x, y) \in \mathcal{X} \times \mathcal{Y}$ is a realization of the random variable (X, Y). The joint distribution $p(x, y)$, and the marginals $p_X(x)$ and $p_Y(y)$, are all unknown. We further denote by b the class imbalance of Y,

$$b = p_Y(1) - p_Y(-1). \tag{1}$$

Let $\{f_i\}_{i=1}^m$ be a set of m binary classifiers operating on $\mathcal{X}$. As our classification problem is binary, the accuracy of the i-th classifier is fully characterized by its sensitivity ψ_i and specificity η_i,

$$\psi_i = \Pr\left(f_i(X) = 1 | Y = 1\right), \ \eta_i = \Pr\left(f_i(X) = -1 | Y = -1\right)$$

For future use, we denote by π_i its balanced accuracy, given by the average of its sensitivity and specificity

$$\pi_i = \tfrac{1}{2}(\psi_i + \eta_i). \tag{2}$$

Note that when the class imbalance is zero, π_i is simply the overall accuracy of the i-th classifier.

The classical conditional independence model. In the model proposed by Dawid and Skene [6], depicted in Fig. 1(left), all m classifiers were assumed conditionally independent given the class label. Namely, for any set of predictions $a_1, \ldots, a_m \in \{\pm 1\}$

$$\Pr(f_1 = a_1, \ldots, f_m = a_m | Y) = \prod_i \Pr(f_i = a_i | Y) \tag{3}$$

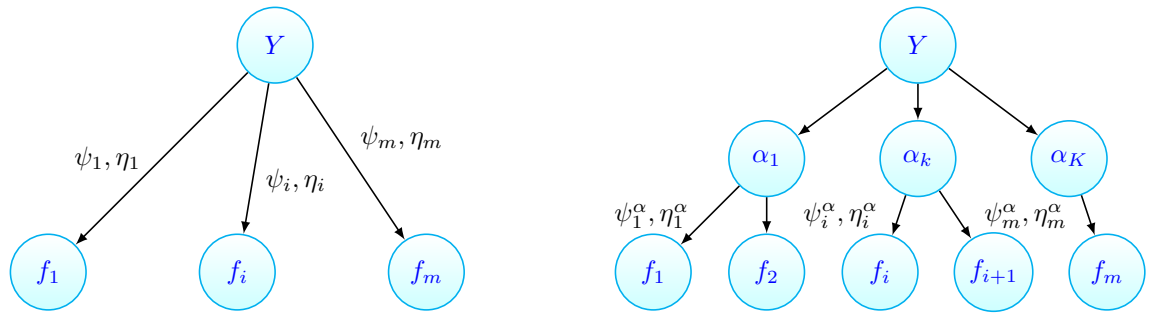

Fig. 1: (Left) The perfect conditional independence model of Dawid and Skene. All classifiers are independent given the class label Y; (Right) The generalized model considered in this work.

As shown in [6], the maximum likelihood estimation (MLE) for y given the parameters ψ_i, η_i and b is *linear* in the predictions of $f_1, ..., f_m$

$$\hat{y} = \text{sign}\,\big(\sum_{i=1}^{m} w_i f_i(x) + w_0\big),\ \ w_i = w(\psi_i, \eta_i). \tag{4}$$

Hence, the main challenge is to estimate the model parameters ψ_i and η_i. A simple approach to do so, as described in [10, 18], is based on the following insight: A classifier which is totally random has zero correlation with any other classifier. In contrast, a high correlation between the predictions of two classifiers is a strong indication that both are highly accurate, assuming they are not both adversarial.

In many realistic scenarios, however, an ensemble may contain several strongly dependent classifiers. Such a scenario has several consequences: First, the above insight that high correlation between two classifiers implies that both are accurate breaks down. Second, as shown in Sec. 5, estimating the classifiers' parameters ψ_i, η_i as if they were conditionally independent may be highly inaccurate. Third, in contrast to Eq. (4), the optimal ensemble learner is in general *non-linear* in the m classifiers. Applying the linear meta-classifier of Eq. (4) may be suboptimal, even when provided with the true classifier accuracies.

A model for conditionally dependent classifiers. In this paper we significantly relax the conditional independence assumption. We introduce a new model which allows classifiers to be dependent through unobserved latent variables, and develop novel methods to learn the model parameters and construct an improved non-linear meta-learner.

In contrast to the 2-layer model of Dawid and Skene, our proposed model, illustrated in Fig. 1(right), has an additional intermediate layer with $K \leq m$ latent binary random variables $\{\alpha_k\}_{k=1}^{K}$. In this model, the unobserved α_k are conditionally independent given the true label Y, whereas each observed classifier depends on Y only through a single and unknown latent variable. Classifiers that depend on different latent variables are thus conditionally independent given Y, whereas classifiers that depend on the same latent variable may have strongly correlated errors. Each hidden variable can be interpreted as a separate unobserved teacher, or source of information, and the classifiers that depend on it are different perturbations of it. Namely, the m observed predictions for each instance are in fact generated by a hidden model with intrinsic dimensionality K, where possibly $K \ll m$.

We now describe in detail our probabilistic model. Since the latent variables $\alpha_1, \ldots, \alpha_K$ are conditionally independent given Y, their distribution is characterized by the class imbalance b and the $2K$ probabilities

$$\Pr(\alpha_k = 1|Y = 1) \quad \text{and} \quad \Pr(\alpha_k = -1|Y = -1).$$

Next, we introduce an assignment function $\mathbf{c} : [m] \to [K]$, such that if classifier f_i depends on α_k then $\mathbf{c}(i) = k$. The dependence of classifier f_i on the class label Y is *only* through its latent variable $\alpha_{\mathbf{c}(i)}$,

$$\Pr(f_i|\alpha_{\mathbf{c}(i)}, Y) = \Pr(f_i|\alpha_{\mathbf{c}(i)}). \tag{5}$$

Hence, classifiers f_i, f_j with $\mathbf{c}(i) \neq \mathbf{c}(j)$ maintain the original conditional independence assumption of Eq. (3). In contrast, classifiers f_i, f_j with $\mathbf{c}(i) = \mathbf{c}(j)$ are only conditionally independent given $\alpha_{\mathbf{c}(i)}$,

$$\begin{aligned}&\Pr(f_i = a_i, f_j = a_j|\alpha_{\mathbf{c}(i)}) = \\ &\Pr(f_i = a_i|\alpha_{\mathbf{c}(i)})\Pr(f_j = a_j|\alpha_{\mathbf{c}(i)}).\end{aligned} \tag{6}$$

Note that if the number of groups K is equal to the number of classifiers, then all classifiers are conditionally independent, and we recover the original model of Dawid and Skene.

Since the model now consists of three layers, the remaining parameters to describe it are the sensitivity

ψ_i^α and specificity η_i^α of the i-th classifier given its latent variable $\alpha_{\mathbf{c}(i)}$,

$$\psi_i^\alpha = \Pr(f_i = 1|\alpha_{\mathbf{c}(i)} = 1),\ \ \eta_i^\alpha = \Pr(f_i = -1|\alpha_{\mathbf{c}(i)} = -1).$$

By Eq. (5), the overall sensitivity ψ_i of the i-th classifier is related to ψ_i^α and η_i^α via

$$\psi_i = \Pr(\alpha_{\mathbf{c}(i)} = 1|Y = 1)\psi_i^\alpha + \Pr(\alpha_{\mathbf{c}(i)} = -1|Y = 1)(1-\eta_i^\alpha) \tag{7}$$

with a similar expression for its overall specificity η_i.

Remark on Model Identifiability. Note that the model depicted in Fig. 1(right) is in general not identifiable. In some cases, the dependency structure can be described by more than one model. For example, the classical model of Dawid and Skene can also be recovered with a single latent variable $K = 1$, by having $\alpha_1 = Y$. See [5] for the exact conditions for identifiability. However, for the purpose of classification and aggregation these models are equivalent, and our purpose is to recover one of them.

Problem Formulation. We consider the following totally unsupervised scenario. Let Z be a binary $m \times n$ matrix with entries $Z_{ij} = f_i(x_j)$, where $f_i(x_j)$ is the label predicted by classifier f_i at instance x_j. We assume x_j are drawn i.i.d. from $p_X(x)$. We also assume the m classifiers satisfy our generalized model, but otherwise we have no prior knowledge as to the number of groups K, the assignment function $\mathbf{c}$ or the classifier accuracies (sensitivities ψ_i,ψ_i^α and specificities η_i, η_i^α). Given only the matrix Z of binary predictions and no labeled data, we consider the following problems:

1. Is it possible to detect strongly dependent classifiers, and estimate the number of groups and the corresponding assignment function $\mathbf{c}$?
2. Given a positive answer to the previous question, how can we estimate the sensitivities and specificities of the m different classifiers and construct an improved, possibly non-linear, meta learner ?

3 Estimating the assignment function

In principle, one could try to estimate all model parameters by maximum likelihood. This results, however, in a hard combinatorial problem. As we show below, the main challenge in our model is to estimate the number of groups K and the assignment function $\mathbf{c}$. Once $\mathbf{c}$ is obtained, we will see in Section 4 that constructing a suitable meta-learner can be reduced to the conditional independent case, already addressed in previous works [10, 11, 18, 28].

We start by estimating the values of K and $\mathbf{c}$. We do so using the low-rank structure of the covariance matrix of the classifiers, implied by our model.

The covariance matrix. Let R denote the $m \times m$ population covariance matrix of the m classifiers

$$r_{ij} = \mathbb{E}[(f_i - \mathbb{E}[f_i])(f_j - \mathbb{E}[f_j])]. \tag{8}$$

The following lemma describes its structure. It generalizes a similar lemma, for the standard Dawid and Skene model, proven in [18]. The proof of this and other lemmas below appear in the appendix.

Lemma 1. *There exists two vectors $v^{on}, v^{off} \in \mathbb{R}^m$ such that for all $i \neq j$,*

$$r_{ij} = \begin{cases} v_i^{off} \cdot v_j^{off} & \text{if } \mathbf{c}(i) \neq \mathbf{c}(j) \\ v_i^{on} \cdot v_j^{on} & \text{if } \mathbf{c}(i) = \mathbf{c}(j) \end{cases} \tag{9}$$

The population covariance matrix is therefore a combination of two rank-one matrices. The block diagonal elements i, j with $\mathbf{c}(i) = \mathbf{c}(j)$ correspond to the rank-one matrix $v^{on}(v^{on})^T$, where on stands for on-block, while the off-block diagonal elements, with $\mathbf{c}(i) \neq \mathbf{c}(j)$ correspond to another rank-one matrix $v^{off}(v^{off})^T$. Let us define the indicator $\mathbb{1}_{\mathbf{c}}(i, j)$

$$\mathbb{1}_{\mathbf{c}}(i, j) = \begin{cases} 1 & \mathbf{c}(i) = \mathbf{c}(j) \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

The non-diagonal elements of R can be written as,

$$r_{ij} = \mathbb{1}_{\mathbf{c}}(i, j) v_i^{on} v_j^{on} + (1 - \mathbb{1}_{\mathbf{c}}(i, j)) v_i^{off} v_j^{off}. \tag{11}$$

Learning the model in the ideal setting. It is instructive to first examine the case where the data is generated according to our model, and the population covariance matrix R is exactly known, i.e. $n = \infty$. The question of interest is how to efficiently recover the assignment function in this setting. To this end, let us look at the possible values of the determinant of 2×2 submatrices of R,

$$M_{ijkl} = \det \begin{pmatrix} r_{ij} & r_{il} \\ r_{kj} & r_{kl} \end{pmatrix} \tag{12}$$

Due to the low rank structure described in Lemma 1, we have the following result, with the exact conditions appearing in the appendix.

Lemma 2. *Assume the two vectors v^{on} and v^{off} are sufficiently different, then $M_{ijkl} = 0$ if and only if either: (i) Three or more of the indices i, j, k and l belong to the same group or (ii) $\mathbf{c}(i) \neq \mathbf{c}(j)$, $\mathbf{c}(j) \neq \mathbf{c}(k)$, $\mathbf{c}(k) \neq \mathbf{c}(l)$ and $\mathbf{c}(l) \neq \mathbf{c}(i)$.*

With details in the appendix, comparing the indices (j, k, l) where $M(i_1, j, k, l) = 0$ with i_1 fixed, to those where $M(i_2, j, k, l) = 0$, we can deduce, in polynomial time, whether $\mathbf{c}(i_1) = \mathbf{c}(i_2)$.

Learning the model in practice. In practical scenarios, the population covariance matrix R is unknown, though we can estimate it by the sample covariance matrix $\hat{R}$. Furthermore, our model is typically only an approximation of the classifiers dependency structure. Given only $\hat{R}$, the approach to recover the assignment function described above, based on exact matching of the pattern of zeros of the determinants of various 2×2 submatrices is not applicable.

In principle, a possible approach to estimate $\mathbf{c}$ from $\hat{R}$, is to define the following residual

$$\Delta(v^{on}, v^{off}, \mathbf{c}) = \sum_{i \neq j} \mathbb{1}_{\mathbf{c}}(i,j)(v_i^{on} v_j^{on} - \hat{r}_{ij})^2 + (1 - \mathbb{1}_{\mathbf{c}}(i,j))(v_i^{off} v_j^{off} - \hat{r}_{ij})^2. \quad (13)$$

and find its global minimum. Unfortunately, as stated in the following lemma, this is not a simple task.

Lemma 3. *Minimizing the residual of Eq.* (13) *for a general covariance matrix $\hat{R}$ is NP-hard.*

In light of Lemma 3, we now present a tractable algorithm to estimate K and $\mathbf{c}$ and provide some theoretical support for it. Our algorithm is inspired by the ideal setting which highlighted the importance of the determinants of 2×2 submatrices. To detect pairs of classifiers f_i, f_j that strongly violate the conditional independence assumption, we thus compute the following score matrix $\hat{S} = \hat{S}(\hat{R})$,

$$\hat{s}_{ij} = \sum_{k \neq i,j; l \neq i,j} |\hat{r}_{ij}\hat{r}_{kl} - \hat{r}_{il}\hat{r}_{kj}|. \quad (14)$$

To motivate Eq. (14), consider the score matrix S computed with the population covariance R. Lemma 2 characterized the cases where the submatrices in Eq. (14) are of rank one, and hence their determinant is zero. When $\mathbf{c}(i) \neq \mathbf{c}(j)$ most submatrices come from four different groups, i.e. will have rank one, and thus the sum s_{ij} will be small. On the other hand, when $\mathbf{c}(i) = \mathbf{c}(j)$ many submatrices will not be rank one and thus s_{ij} will be large, assuming no degeneracy between v^{on} and v^{off}. As $\hat{S} \xrightarrow{n \to \infty} S$, large values of $\hat{s}_{ij}$ indicate a strong conditional dependence between classifiers f_i and f_j.

The following lemma provides some theoretical justification for the utility of the score matrix S computed with the population covariance, in recovering the assignment function $\mathbf{c}$. For simplicity, we analyze a symmetric case where the class imbalance $b = 0$, $\Pr(\alpha_k = -1 | y = -1) = \Pr(\alpha_k = 1 | y = 1)$ and all groups have equal size of m/K. We measure deviation from conditional independence by the following

Algorithm 1 Estimating the assignment function $\mathbf{c}$ and vectors v^{on}, v^{off}

1: Estimate the covariance matrix R (8).
2: Obtain the score matrix by (14)
3: **for all** $1 < k < m$ **do**
4: Estimate $\mathbf{c}$ by performing spectral clustering with the Laplacian of the score matrix.
5: Use the clustering function to estimate the two vectors v^{on}, v^{off}.
6: Calculate residual by (13).
7: **end for**
8: Pick the assignment function and vectors which yield minimal residual.

conditional covariance matrices, C^+ and C^-,

$$\begin{aligned} c_{ij}^+ &= \mathbb{E}[(f_i - \mathbb{E}[f_i])(f_j - \mathbb{E}[f_j]) | Y = 1] \\ c_{ij}^- &= \mathbb{E}[(f_i - \mathbb{E}[f_i])(f_j - \mathbb{E}[f_j]) | Y = -1]. \quad (15) \end{aligned}$$

Finally, we assume the balanced accuracies of all classifiers satisfy $(2\pi_i - 1) > \delta$ for some $\delta > 0$.

Lemma 4. *Under the assumptions described above, if $\mathbf{c}(i) = \mathbf{c}(j)$ then*

$$s_{ij} > m^2 \left(1 - \frac{3}{K}\right) \delta^2 |c_{ij}^+| = m^2 \left(1 - \frac{3}{K}\right) \delta^2 |c_{ij}^-| \quad (16)$$

In contrast, if $\mathbf{c}(i) \neq \mathbf{c}(j)$ then

$$s_{ij} < \frac{m^2}{K} \left(5 - \frac{8}{K}\right). \quad (17)$$

An immediate corollary from lemma 4, is that if the classifiers are sufficiently accurate, and their dependencies within each group are strong enough then the score matrix exhibits a clear gap with $\max_{\mathbf{c}(i) \neq \mathbf{c}(j)} S_{ij} < \min_{\mathbf{c}(i) = \mathbf{c}(j)} S_{ij}$. In this case, even a simple single-linkage hierarchical clustering algorithm can recover the correct assignment function from S. In reality, as only $\hat{S}$ is available, we apply spectral clustering which is more robust, and works better in practice.

We illustrate the usefulness of the score matrix using the DREAM challenge S1 dataset, which contains $m = 124$ classifiers. Fig. 2 shows the matrix of conditional covariance $\frac{1}{2}(C^+ + C^-)$ of Eq. (15), computed using the ground truth labels. Fig. 3 shows the score matrix $\hat{S}$ computed using only the classifiers predictions. We also plot the values of the score matrix vs. the conditional covariance in figure 4. Clearly, a high score is a reliable indication for strong conditional dependencies between classifiers.

It is important to note that the time complexity to build the score matrix S is $\mathcal{O}(m^4)$, which in general

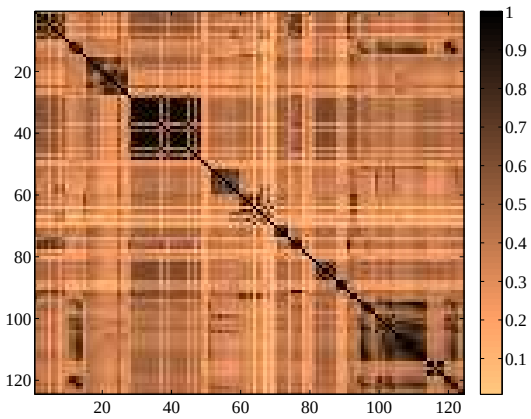

Fig. 2: The *conditional* covariance matrix $\frac{1}{2}(C^+ + C^-)$ of the DREAM dataset S1, computed using the ground truth labels.

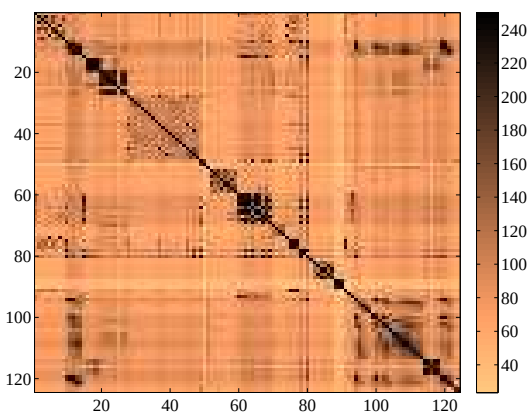

Fig. 3: The score matrix $\hat{S}$ of the DREAM S1 dataset, computed from the matrix of classifier predictions. For visualization purposes, the upper limit of the above score matrix is fixed at 250.

may be considered too expensive. However, in many real world problems where the number of classifiers $m \approx 100$, our algorithm typically runs in less than an hour. If needed, this can be sped-up by sampling elements of S instead of computing the full matrix [9].

Estimating the assignment function c. We estimate $\mathbf{c}$ by spectral clustering the score matrix $\hat{S}$ of Eq. (14). As the number of clusters or groups K is unknown, we choose the one which minimizes the residual function defined in Eq. (13). The steps for estimating the number of groups K and the assignment function $\mathbf{c}$ are summarized in Algorithm 1. Note that retrieving v^{on} and v^{off} from the covariance matrix is a rank-one matrix completion problem, for which several solutions exist, for example see [4]. Also note that while we perform spectral clustering for various number of clusters, the costly eigen-decoposition step only needs to be done once.

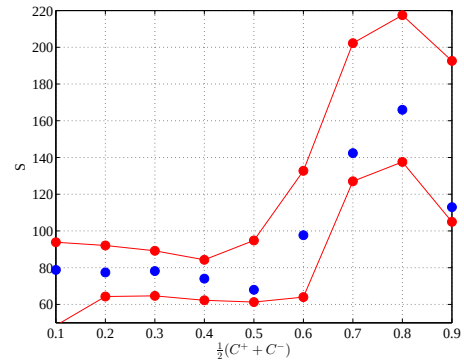

Fig. 4: Values of S vs. the corresponding conditional covariance matrix $\frac{1}{2}(C^+ + C^-)$ for the DREAM dataset S1. The blue dots represent the mean value, the upper and lower red dots represent the $80th$ and $20th$ quantiles, respectively.

4 The latent spectral meta learner

Estimating the model parameters. Given estimates of K and of the assignment function $\mathbf{c}$, estimating the remaining model parameters can be divided into two stages: (i) Estimating the sensitivity and specificity of the different classifiers given the latent variables α_k: $\psi_i^\alpha, \eta_i^\alpha$ (ii) Estimating the probabilities associated with the latent variables, $\Pr(\alpha_k = 1|Y = 1)$ and $\Pr(\alpha_k = -1|Y = -1)$.

The key observation is that each of these stages can be solved by the classical conditional independent model of [6]. In particular, classifiers with a common latent variable are conditionally independent given its value. Similarly the K latent variables themselves are conditionally independent given the true label Y. Thus, we can solve the two stages sequentially by any of the various methods already developed for the Dawid and Skene model. In our implementation, we used the spectral meta learner proposed in [10], whose code is publicly available. A pseudo-code for this process appears in Algorithm 2.

Label Predictions. We predict the label of an instance x by maximing the likelihood function

$$\hat{y} = \operatorname*{argmax}_{y=\pm 1} \Pr(f_1(x), \ldots, f_m(x)|y) \tag{18}$$

that corresponds to our generative model, Fig. 1(right). The above probability depends on the model parameters $b, \psi_i^\alpha, \eta_i^\alpha, \psi_\alpha, \eta_\alpha$, and the assignment function $\mathbf{c}$ for which we use the values estimated above.

Classifier selection. In some cases, it is required to construct a *sparse* ensemble learner which uses only a small subset of at most M out of the available m

Algorithm 2 Estimate model parameters

1: **Input:** Matrix of predictions $f_i(x_j)$, parameters K and $\mathbf{c}$.
2: **for** $k = 1, .., K$ **do**
3: Find all classifiers f_i where $\mathbf{c}(i) = k$
4: Estimate $\psi_i^\alpha, \eta_i^\alpha$ and $\mathbb{E}[\alpha_k]$
5: Estimate the latent values $\alpha_k(x_j)$, $\forall j = 1, ..., n$
6: **end for**
7: Estimate $\Pr(\alpha_k = 1|Y = 1), \Pr(\alpha_k = -1|Y = -1)$

classifiers. This problem of selecting a small subset of classifiers, known as *ensemble pruning*, has mostly been studied in supervised settings, see [16, 22, 27].

Under the conditional independence assumption, the best subset simply consists of the M most accurate classifiers. In our model, in contrast, the correlations between the classifiers have to be taken into account. Assuming the required number of classifiers is smaller than the number of groups $M \leq K$, a simple approach is to select the M most accurate classifiers under the constraint that they all come from different groups. This creates a balance between accuracy and diversity.

5 Experiments

We demonstrate the performance of the latent variable model on artificial data, on datasets from the UCI repository and on the ICGA-TCGA dream challenge. Code is available at the author's website.

We compared the following unsupervised ensemble methods: (1) `Majority voting`, which serves as a baseline; (2) `SML+EM` - a spectral meta-learner based on the independence assumption [10] providing an initial guess followed by EM iterations; (3) `Oracle-CI`: A linear meta-learner based on Eq. (4), which assumes conditional independence but is given the exact accuracies of all the individual classifiers. (4) `L-SML` (latent SML), the new algorithm presented in this work.

On the artificial data, we also compute its oracle meta-learner, denoted `Oracle-L`, which is given the exact structure and parameters of the model, and predicts the label Y by maximum likelihood. For the UCI datasets we also compared to CUBAM (Caltech UCSD Binary Annotation Model) [25]. For the UCI and Dream challenge, we added a comparison to a supervised method - a linear SVM classifier trained with the predictions of the m classifiers as its input features.

5.1 Artificial Data

To validate our theoretical analysis, we generated artificial binary data according to our assumed model,

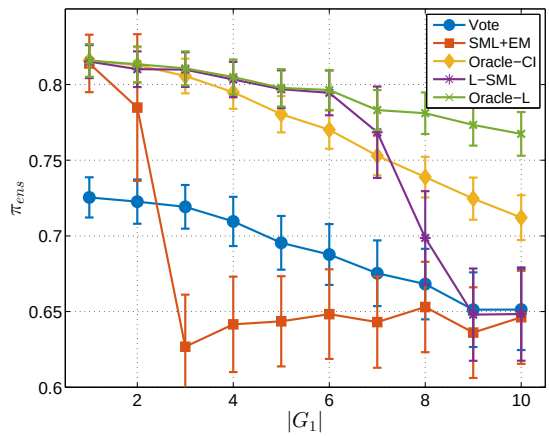

Fig. 5: Simulated data: Ensemble learner balanced accuracy vs. the size of group 1.

on a balanced classification problem with $b = 0$. We generated an ensemble of $m = 20$ classifiers with $n = 10^4$ instances. We consider the case where there is only one group G_1 of correlated classifiers, with the remaining $m - |G_1|$ classifiers all conditionally independent. All the parameters of the ensemble were chosen uniformly at random from the following intervals: $\Pr(\alpha = 1|Y = 1), \Pr(\alpha = -1|Y = -1) \in [0.5, 0.8], \{\psi_i^\alpha, \eta_i^\alpha\} \in [0.7, 0.9]$. Note that for $|G_1| = 1$ all classifiers are conditionally independent. Fig. 5 compares the balanced accuracy of the five unsupervised ensemble learners described above, as a function of the size of the first group $|G_1|$. As seen in Fig. 5, up to $|G_1| = 7$, the ensemble learner based on the concept of correlated classifiers achieves similar results to the optimal classifier ('oracle-L'). As expected from Lemma 4, as $|G_1|$ increases, it is harder to correctly estimate $\mathbf{c}$ with the score matrix. A figure with a similar experiment with $m = 40$ classifiers appears in the appendix. There, our algorithm maintained its stability up to $|G_1|$=16. This is in accordance with Lemma 4, since the parameter that determines the performance of our algorithm is not the absolute size of G_1, but the ratio between its size and the number of classifiers.

A complementary figure showing the probability to recover the correct assignment function as a function of $|G_1|$ appears in the appendix. As expected, the degradation in performance starts when the algorithm fails to correctly estimate the model structure.

5.2 UCI data sets

For each of the following 4 UCI datasets, Magic, Spambase, Miniboo and Musk, we constructed an ensemble of $m = 16$ classifiers, consisting of: 4 random forests, 3 logistic model trees, 4 SVM and 5 naive Bayes. Each classifier was trained on a separate, randomly chosen

labeled dataset. In our unsupervised ensemble scenario we had access only to their predictions on a large independent test set.

We present results for the magic dataset, which contains 19000 instances with 11 attributes. The task is to classify each instance as background or high energy gamma rays. As seen in Fig. 6, the L-SML improves substantially over the standard SML, and even on the oracle classifier that assumes conditional independence. Our method also outperforms the best individual classifier as well as the supervised method.

Figs. 10 and 11 in the appendix show the conditional covariance matrix and our assignment. It can be observed that strongly dependent classifiers are indeed grouped together correctly. Fig. 16 compares the classifiers error rate estimated by the L-SML and by the method presented in [19]. For additional results on the other UCI datasets, see the appendix.

5.3 The DREAM mutation calling challenge

The ICGC-TCGA DREAM challenge is an international effort to improve standard methods for identifying cancer-associated mutations and rearrangements in whole-genome sequencing (WGS) data. This publicly available database contains both real and synthetic in-silico tumor instances. The database contains 14 different datasets, each with over 100,000 instances.

Participants in the competition are given access to the predictions of about a hundred different classifiers (denoted there as pipe-lines)[1]. These classifiers were constructed by various labs worldwide, each employing their own biological knowledge and possibly proprietary labeled data. The two current challenges are to construct a meta-learner, by using either (1) all m classifiers; or (2) at most five of them. We evaluated proposed meta-classifiers f_{mc} by their balanced error,

$$1-\pi = \frac{1}{2}(\Pr(f_{mc}=1|y=-1)+\Pr(f_{mc}=-1|y=1)).$$

Below we present results on the datasets S1, S2 and S3 for which ground-truth labels have been released.

	Mean	Best	Vote	SML EM	Or-CI	L-SML	SVM
1	6.1	1.7	2.8	1.7	1.7	**1.6**	1.82
2	8.7	1.8	4.0	2.8	2.8	**2.3**	1.23
3	8.3	2.5	4.3	2.3	2.3	**1.8**	1.7

Table 1: DREAM challenge I. Balanced error using full ensemble, the first two columns give the mean and smallest balanced error of all classifiers.

[1]The data is available at http://dreamchallenges.org/

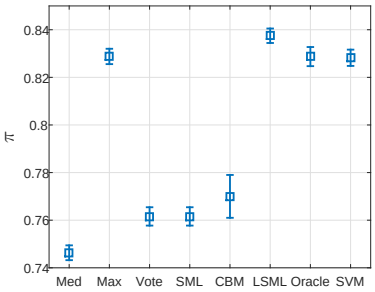

Fig. 6: UCI magic dataset, a comparison of four unsupervised ensemble learners, alongside the oracle ensemble based on the conditional independence assumption, and a SVM supervised ensemble.

	Vote	SML+EM	Oracle-CI	L-SML
S1	3.2	2.3	**1.9**	2.0
S2	4.3	4.1	**2.5**	2.8
S3	2.9	2.9	2.8	**2.5**

Table 2: Balanced error of sparse meta-classifiers.

Challenge I. The balanced errors of the different meta-learners, constructed using all m classifiers, are given in table 1. The L-SML method outperforms the other meta-learners in all three datasets. On dataset S3 it reduces the balanced error by more than 20% over competing meta learners.

Challenge II. Here the goal is to construct a sparse meta-learner based on at most five individual classifiers from the ensemble. For the methods based on the Dawid and Skene model (SML+EM, voting and Oracle-CI), we took the 5 classifiers with the highest estimated (or known) balanced accuracies. For our model, since the estimated number of groups is larger than five, we took the five classifiers with highest estimated balanced accuracies, under the constraint that they belong to five different groups. For all methods, the final prediction was made by a simple vote of the five chosen classifiers. Though potentially suboptimal, we nonetheless use it as our purpose was to compare the diversity of the chosen classifiers. As seen in table 2, our method outperforms voting and SML, and is comparable to the oracle learner.

Acknowledgments. We thank Anthony Platanios for interesting discussions and for providing us with his code. This research was funded in part by the Intel Collaborative Research Institute for Computational Intelligence (ICRI-CI). Y.K. is supported by the National Institutes of Health Grant R0-1 CA158167, and R0-1 GM086852. This work was also partially supported by NIH grant 1R01HG008383-01A1 (Y.K. and B.N.).

References

[1] N. Aghaeepour, G. Finak, H. Hoos, T.R Mosmann, R Brinkman, R Gottardo, R H Scheuermann, FlowCAP Consortium, and DREAM Consortium. Critical assessment of automated flow cytometry data analysis techniques. *Nature methods*, 10(3):228–238, 2013.

[2] A. Anandkumar, R. Ge, D. Hsu, S.M. Kakade, and M. Telgarsky. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15:2773–2832, 2014.

[3] PC Boutros, AA Margolin, JM Stuart, A Califano, and G Stolovitzky. Toward better benchmarking: challenge-based methods assessment in cancer genomics. *Genome biology*, 15(9):462, 2014.

[4] E.J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9:717–772, 2009.

[5] J.T Chang. Full reconstruction of markov models on evolutionary trees: Identifiability and consistency. *Mathematical biosciences*, 137(1):51–73, 1996.

[6] A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the em algorith. *Journal of the Royal Statistical Society. Series C*, 28:20–28, 1979.

[7] P. Donmez, G. Lebanon, and K. Balasubramanian. Unsupervised supervised learning i: Estimating classification and regression errors without labels. *Journal of Machine Learning Research*, 11:1323–1351, 2010.

[8] A.D. Ewing, K.E. Houlahan, Y. Hu, K. Ellrott, C. Caloian, T.N. Yamaguchi, J.Ch. Bare, C. P'ng, D. Waggott, and V.Y. Sabelnykova. Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. *Nature methods*, 12:623–630, 2015.

[9] E. Fetaya, O. Shamir, and S. Ullman. Graph approximation and clustering on a budget. *18th conference on artificial intelligence and statistics*, 2015.

[10] A. Jaffe, B. Nadler, and Y. Kluger. Estimating the accuracies of multiple classifiers without labeled data. In *18th conference on artificial intelligence and statistics*, pages 407–415, 2015.

[11] P. Jain and S. Oh. Learning mixtures of discrete product distributions using spectral decompositions. *Journal of Machine Learning Research*, 35:1–33, 2014.

[12] Valen E Johnson. On bayesian analysis of multirater ordinal data: An application to automated essay grading. *Journal of the American Statistical Association*, 91(433):42–51, 1996.

[13] Valen E Johnson, Robert O Deaner, and Carel P Van Schaik. Bayesian analysis of rank data with application to primate intelligence experiments. *Journal of the American Statistical Association*, 97(457):8–17, 2002.

[14] D.R. Karger, S. Oh, and D. Shah. Budget-optimal crowdsourcing using low-rank matrix approximations. In *IEEE Alerton Conference on Communication, Control and Computing*, pages 284–291, 2011.

[15] John A Lee. Click to cure. *The Lancet Oncology*, 2013.

[16] G. Martinez-Muoz, D. Hernández-Lobato, and A. Suarez. An analysis of ensemble pruning techniques based on ordered aggregation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(2):245–259, 2009.

[17] M Micsinai, F Parisi, F Strino, P Asp, B.D Dynlacht, and Y Kluger. Picking chip-seq peak detectors for analyzing chromatin modification experiments. *Nucleic acids research*, 2012.

[18] F. Parisi, F. Strino, B. Nadler, and Y. Kluger. Ranking and combining multiple predictors without labeled data. *Proceedings of the National Academy of Sciences*, 111:1253–1258, 2014.

[19] E.A. Platanios, A. Blum, and T. Mitchell. Estimating accuracy from unlabeled data. In *Uncertainty in Artificial Intelligence*, 2014.

[20] Alexander James Quinn. Crowdsourcing decision support: frugal human computation for efficient decision input acquisition. *PhD thesis*, 2014.

[21] V.C. Raykar, Y. Shipeng, L.H. Zhao, G.H. Valdez, C. Florin, L. Bogoni, and Moy L. Learning from crowds. *J. Machine Learning Research*, 11:1297–1322, 2010.

[22] L. Rokach. Collective-agreement-based pruning of ensembles. *Computational Statistics & Data Analysis*, 53(4):1015–1026, 2009.

[23] A. Sheshadri and M. Lease. Square: A benchmark for research on computing crowd concensus. In *AAAI conference on human computation and crowdsourcing*, 2013.

[24] Tian Tian and Jun Zhu. Uncovering the latent structures of crowd labeling. In *Advances in Knowledge Discovery and Data Mining*, pages 392–404. Springer, 2015.

[25] P. Welinder, S. Branson, S. Belongie, and P. Perona. The multidimensional wisdom of crowds. In *Advances in Neural Information Processing Systems 23 (NIPS 2010)*, 2010.

[26] J. Whitehill, P. Ruvolo, T. Wu, J Bergsma, and J.R. Movellan. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in Neural Information Processing Systems 22 (NIPS 2009)*, 2009.

[27] Xu-Cheng Yin, Kaizhu Huang, Chun Yang, and Hong-Wei Hao. Convex ensemble learning with sparsity and diversity. *Information Fusion*, 2014.

[28] Y. Zhang, X. Chen, D. Zhou, and M.I. Jordan. Spectral methods meet EM: A provably optimal algorithm for crowdsourcing. In *Advances in Neural Information Processing Systems*, volume 27, pages 1260–1268, 2014.