A Proof of Lemma 1

This proof is based on the following lemma, which appears in [18]:
If two classifiers f_i, f_j are conditionally independent given the class label Y, then the covariance between them is equal to

$$r_{ij} = (1-b^2)(\psi_i + \eta_i - 1)(\psi_j + \eta_j - 1). \tag{19}$$

In our model, if $\mathbf{c}(i) \neq \mathbf{c}(j)$, then f_i, f_j are indeed conditionally independent (Fig. 1,right). The first part of lemma 1 thus follows directly from Eq. (19), with $v_i^{off} = \sqrt{1-b^2}(\psi_i + \eta_i - 1)$.

To prove the second part of lemma 1, we note that according to our model, two classifiers f_i, f_j with $\mathbf{c}(i) = \mathbf{c}(j)$ are conditionally independent given the value of their latent variable α. Therefore, we can treat α as the class label, and apply Eq. (19) with b replaced by the expectation of α, and the sensitivity and specificity ψ_i, η_i replaced by $\psi_i^\alpha, \eta_i^\alpha$ respectively. Hence, Eq. (19) becomes,

$$r_{ij} = (1-\mathbb{E}[\alpha]^2)(\psi_i^\alpha + \eta_i^\alpha - 1)(\psi_j^\alpha + \eta_j^\alpha - 1) = v_i^{on} v_j^{on}, \tag{20}$$

where $v_i^{on} = \sqrt{1-\mathbb{E}[\alpha]^2}(\psi_i^\alpha + \eta_i^\alpha - 1)$, and $\alpha = \alpha_{\mathbf{c}(i)}$.

B Proof of Lemma 2

We assume that v^{on} and v^{off} are sufficiently different in the following precise sense: We require that for all 4 distinct indices i, j, k, l, $v_i^{on} \cdot v_j^{on} \cdot v_k^{on} \cdot v_l^{on} \neq v_i^{off} \cdot v_j^{off} \cdot v_k^{off} \cdot v_l^{off}$.

Next, we elaborate on the relation between v_i^{off} and v_i^{on}. Let us denote by $\psi_\alpha^y, \eta_\alpha^y$ the sensitivity and specificity of the latent variable α. Let f_i be a classifier that depends on α. Applying Bayes rule, its overall sensitivity and specificity are given by,

$$\psi_i = \psi_\alpha^y \psi_i^\alpha + (1-\psi_\alpha^y)(1-\eta_i^\alpha) \qquad \eta_i = \eta_\alpha^y \eta_i^\alpha + (1-\eta_\alpha^y)(1-\psi_i^\alpha).$$

Adding ψ_i and η_i we get the following,

$$\psi_i + \eta_i - 1 = (\psi_\alpha^y + \eta_\alpha^y - 1)(\psi_i^\alpha + \eta_i^\alpha - 1). \tag{21}$$

If $\mathbf{c}(i) = \mathbf{c}(j)$ we have the following dependency between (v_i^{off}, v_j^{off}) and (v_i^{on}, v_j^{on}),

$$\left[\begin{array}{c} v_i^{off} \\ v_j^{off} \end{array}\right] = (1-b^2)(\psi_\alpha^i + \eta_\alpha^i - 1)\left[\begin{array}{c} (\psi_i^\alpha + \eta_i^\alpha - 1) \\ (\psi_j^\alpha + \eta_j^\alpha - 1) \end{array}\right] = \frac{(1-b^2)(\psi_\alpha^i + \eta_\alpha^i - 1)}{\sqrt{1-\mathbb{E}[\alpha]^2}}\left[\begin{array}{c} v_i^{on} \\ v_j^{on} \end{array}\right] \tag{22}$$

It follows that two elements v_i^{off}, v_j^{off} where $\mathbf{c}(i) = \mathbf{c}(j)$ are linearly dependent with the corresponding elements of v_i^{on}, v_j^{on}. This fact shall be useful in proving the lemma.

To prove lemma 2 we analyze all various possibilities for the group assignments of the four indices i, j, k, l of

$$M(i,j,k,l) = \det\left(\begin{array}{cc} r_{ij} & r_{il} \\ r_{kj} & r_{kl} \end{array}\right)$$

.

1. $c(i) = c(j) = c(k) = c(l)$: In this case $M(i,j,k,l) = v_i^{on} v_j^{on} v_k^{on} v_l^{on} - v_i^{on} v_l^{on} v_k^{on} v_j^{on} = 0$.
2. $c(i) \neq c(j)$, and $c(j) \neq c(k)$, and $c(k) \neq c(l)$ and $c(l) \neq c(i)$: Here $M(i,j,k,l) = v_i^{off} v_j^{off} v_k^{off} v_l^{off} - v_i^{off} v_l^{off} v_k^{off} v_j^{off} = 0$.
3. $c(i) = c(l) = c(k) \neq c(j)$: $M(i,j,k,l) = v_i^{off} v_j^{off} v_k^{on} v_l^{on} - v_i^{on} v_l^{on} v_k^{off} v_j^{off} = v_j^{off} v_l^{on}\left(v_i^{off} v_k^{on} - v_i^{on} v_k^{off}\right)$. From the linear dependency shown in Eq. (22), $\left(v_i^{off} v_k^{on} - v_i^{on} v_k^{off}\right) = 0$.
4. $c(i) = c(j)$, $c(k) = c(l)$ and $c(i) \neq c(k)$: $M(i,j,k,l) = v_i^{on} v_j^{on} v_k^{on} v_l^{on} - v_i^{off} v_l^{off} v_k^{off} v_j^{off} \neq 0$ from our assumption.

It can be seen that M_{ijkl} is equal to zero *only* if either three or more of the indices are equal (cases (1) and (2)) or all four pairs of indices which appear in the determinant belong to different groups (case (3)).

C Algorithm for the ideal setting

An immediate conclusion from lemma 2, is that the indices i, j, k and l for which $M(i,j,k,l) = 0$ depend only on the assignment function. This means we can compare the pattern of zeros for $M(i_1,j,k,l)$ and $M(i_2,j,k,l)$ to decide if f_{i1} and f_{i2} belong to the same group. If $c(i_1) = c(i_2)$ then $M(i_1,j,k,l) = 0 \iff M(i_2,j,k,l) = 0$. On the other hand if $c(i_1) \neq c(i_2)$ and at least one of the indices i_1 and i_2 , w.l.o.g i_1, belongs to a group with more than one element, then there exist indices j, k and l such that $M(i_1,j,k,l) \neq 0$ but $M(i_2,j,k,l) = 0$. This occurs when $c(i_1) = c(j)$, and $c(i_2) \neq c(j) \neq c(k) \neq c(l)$.

This means that by comparing the pattern of zeros, we can recover the assignment function. Notice, that according to the algorithm, all singleton classifiers, that is, classifiers who are conditionally independent with the rest of the ensemble, are grouped together under a common latent variable. This is not a problem, as our model is not unique and this is an equivalent probabilistic model, with the associated latent variable being identical to Y.

Algorithm 3 Check if $\mathbf{c}(i_1) = \mathbf{c}(i_2)$

Initialize $(m-2) \times (m-3) \times (m-4)$ arrays T_1, T_2 to zero
for $j \neq k \neq l \neq i_1, i_2$ **do**
    if $r_{i_1 j} r_{kl} - r_{i_1 l} r_{kj} = 0$ **then** $(T_1(j,k,l) = 1)$
    end if
    if $r_{i_2 j} r_{kl} - r_{i_2 l} r_{kj} = 0$ **then** $(T_2(j,k,l) = 1)$
    end if
end for
if $(T_1 = T_2)$ **then**
    $\mathbf{c}(i_1) = \mathbf{c}(i_2)$.
else
    $\mathbf{c}(i_1) \neq \mathbf{c}(i_2)$.
end if

D Minimizing Δ is a NP hard problem

We prove lemma 3 for the case of $K = 2$ clusters and known vectors $\boldsymbol{v}^{off}, \boldsymbol{v}^{on}$. Our goal is to find a minimizer for the following residual:

$$\hat{\mathbf{c}} = \operatorname*{argmin}_{\mathbf{c}} \Delta(\mathbf{c}) = \operatorname*{argmin}_{\mathbf{c}} \sum_{i,j} \mathbb{1}_{\mathbf{c}}(i,j)(v_i^{on} v_j^{on} - r_{ij})^2 + (1 - \mathbb{1}_{\mathbf{c}}(i,j))(v_i^{off} v_j^{off} - r_{ij})^2 \tag{23}$$

For the case of $K = 2$ we can simplify the residual considerably. Let us define a vector $\boldsymbol{x} \in \{-1,1\}^m$ where $x_i = 1$ if $\mathbf{c}(i) = 1$ and $x_i = -1$ if $\mathbf{c}(i) = 2$. We can replace the indicator function $\mathbb{1}(i,j)$ with the following,

$$\mathbb{1}(i,j) = \frac{(1 + x_i x_j)}{2}, \qquad 1 - \mathbb{1}(i,j) = \frac{(1 - x_i x_j)}{2}. \tag{24}$$

In addition, we can replace the minimization over $\mathbf{c}$ with a minimization over $\boldsymbol{x}$,

$$\begin{aligned} \hat{\boldsymbol{x}} &= \operatorname*{argmin}_{\boldsymbol{x} \in \{\pm 1\}^m} \sum_{i,j} \frac{(1 + x_i x_j)}{2}(v_i^{on} v_j^{on} - r_{ij})^2 + \frac{(1 - x_i x_j)}{2}(v_i^{off} v_j^{off} - r_{ij})^2 \\ &= \operatorname*{argmin}_{\boldsymbol{x} \in \{\pm 1\}^m} \sum_{i,j} \frac{1}{2}\left((v_i^{on} v_j^{on} - r_{ij})^2 + (v_i^{off} v_j^{off} - r_{ij})^2\right) \\ &\quad + \frac{x_i x_j}{2}\left((v_i^{on} v_j^{on} - r_{ij})^2 + (v_i^{off} v_j^{off} - r_{ij})^2\right). \end{aligned} \tag{25}$$

The first term does not depend on $\boldsymbol{x}$ and we can omit it from the minimization problem. Let us also define the matrix $\tilde{R}$,

$$\tilde{r}_{ij} = \frac{\left((v_i^{on} v_j^{on} - r_{ij})^2 + (v_i^{off} v_j^{off} - r_{ij})^2\right)}{2} \tag{26}$$

We are left with the following minimization problem:

$$\hat{\boldsymbol{x}} = \underset{\boldsymbol{x}\in\{\pm1\}^m}{\text{argmin}} \sum_{i,j} x_i x_j \tilde{r}_{ij} = \underset{\boldsymbol{x}\in\{\pm1\}^m}{\text{argmin}}\ \boldsymbol{x}^T \tilde{R}\boldsymbol{x} \tag{27}$$

If there is a binary vector whose residual is precisely zero, then it can be found by computing the eigenvector with smallest eigenvalue of the matrix $\tilde{R}$. If, however, the minimal residual is not zero, then eq. (27) is a quadratic optimization problem involving discrete variables, which is well known to be a NP-hard problem.

E Proof of Lemma 4

We start by proving the first part of the lemma, where $\mathbf{c}(i) = \mathbf{c}(j)$. The score matrix s_{ij} is a sum of all possible 2×2 determinants,

$$s_{i,j} = \sum_{k,l\neq i,j} |r_{ij}r_{kl} - r_{il}r_{jk}| = \sum_{k,l\neq i,j} s_{ij}^{kl}, \tag{28}$$

where we define s_{ij}^{kl} as a single score element. The following table separates the various score elements s_{ij}^{kl} into four types, and states the number of elements in each type.

Element type	Number of elements
$\mathbf{c}(i) = \mathbf{c}(j) \neq \mathbf{c}(k) \neq \mathbf{c}(l)$	$m^2\left(1-\frac{3}{K}+\frac{2}{K^2}\right)$
$\mathbf{c}(i) = \mathbf{c}(j) \neq \mathbf{c}(k) = \mathbf{c}(l)$	$m^2\left(1-\frac{1}{K}\right)\left(\frac{1}{K}-\frac{1}{m}\right)$
$\mathbf{c}(i) = \mathbf{c}(j) = \mathbf{c}(k) \neq \mathbf{c}(l)$	$2m^2\left(\frac{1}{K}-\frac{2}{m}\right)\left(1-\frac{1}{K}\right)$
$\mathbf{c}(i) = \mathbf{c}(j) = \mathbf{c}(k) = \mathbf{c}(l)$	$m^2\left(\frac{1}{K}-\frac{2}{m}\right)\left(\frac{1}{K}-\frac{3}{m}\right)$

According to lemma 1, the contribution to the score from elements of the third and fourth type is exactly 0 (see details in Sec. B). We will focus on analyzing the score elements of the first type $\mathbf{c}(i) = \mathbf{c}(j) \neq \mathbf{c}(k) \neq \mathbf{c}(l)$, which is the dominant factor assuming $K \geq 4$. Let us denote by π_i^α the balanced accuracy of classifier i with relation to $\alpha_{\mathbf{c}(i)}$,

$$\pi_i^\alpha = \frac{1}{2}(\psi_i^\alpha + \eta_i^\alpha)$$

Recall, that we assume a symmetrical case where $b = 0$, and $\Pr(\alpha = 1|y = 1) = \Pr(\alpha = -1|y = -1)$. These assumptions imply that $\mathbb{E}[\alpha_k] = 0$ for all $k = 1...K$.

Let us consider Lemma 1 in order to analyze the value of s_{ij}^{kl},

$$\begin{aligned} s_{ij}^{kl} &= |r_{ij}r_{kl} - r_{ij}r_{jk}| = |(2\pi_i^\alpha - 1)(2\pi_j^\alpha - 1)(2\pi_k - 1)(2\pi_l - 1) - (2\pi_i - 1)(2\pi_j - 1)(2\pi_k - 1)(2\pi_l - 1)| \\ &= |(2\pi_k - 1)(2\pi_l - 1)\big((2\pi_i^\alpha - 1)(2\pi_j^\alpha - 1) - (2\pi_i - 1)(2\pi_j - 1)\big)| \end{aligned} \tag{29}$$

For simplicity of notation, let us denote by γ the ratio of true positives and negatives of the latent variables:

$$\gamma = \Pr(\alpha_k = 1|Y = 1) = \Pr(\alpha_k = -1|Y = -1) \tag{30}$$

It can easily be shown that the following holds:

$$(2\pi_i - 1) = (2\gamma - 1)(2\pi_i^\alpha - 1) \qquad (2\pi_j - 1) = (2\gamma - 1)(2\pi_j^\alpha - 1) \tag{31}$$

Inserting (31) into (29) we get,

$$\begin{aligned} s_{ij}^{kl} = |(2\pi_k - 1)(2\pi_l - 1)(2\pi_i^\alpha - 1)(2\pi_j^\alpha - 1)(1 - (2\gamma - 1)^2)| = \\ |4(2\pi_k - 1)(2\pi_l - 1)(2\pi_i^\alpha - 1)(2\pi_j^\alpha - 1)(\gamma(1-\gamma))| \end{aligned} \tag{32}$$

Let us now derive the values of the conditional covariance matrices C^+, C^-. First, In order to obtain C^+, we can apply the first part of Lemma 1,

$$C^+ = \mathbf{E}[(f_i - \mu_i)(f_j - \mu_j)|Y = 1] = (1 - \mathbb{E}[\alpha_{\mathbf{c}(i)}|Y = 1]^2)(2\pi_i^\alpha - 1)(2\pi_j^\alpha - 1) \tag{33}$$

A similar argument applies to C^-, with $E[\alpha_{\mathbf{c}(i)}|Y=-1]$. The conditional expectation of α is equal to,

$$\mathbb{E}[\alpha|Y=1] = 2\gamma - 1 \quad \mathbb{E}[\alpha|Y=-1] = 1-2\gamma \tag{34}$$

A simple derivation yields the following for both cases,

$$(1-\mathbb{E}[\alpha|Y=1]^2) = (1-\mathbb{E}[\alpha|Y=-1]^2) = 4\gamma(1-\gamma) \tag{35}$$

The value of c^+_{ij} is therefore equal to c^-_{ij}, and both are equal to the following,

$$c^+_{ij} = c^-_{ij} = 4\gamma(1-\gamma)(2\pi^\alpha_i - 1)(2\pi^\alpha_j - 1) \tag{36}$$

Inserting (36) into (32) yields,

$$s^{kl}_{ij} = |(2\pi_k - 1)(2\pi_l - 1)c^+_{ij}| = |(2\pi_k - 1)(2\pi_l - 1)c^-_{ij}| \tag{37}$$

For simplicity, since $C^+ = C^-$, we will use only C^+. The total score contribution of the first type of elements is therefore,

$$\sum_{k,l} s^{kl}_{ij} = |c^+_{ij}| \sum_{k,l} |(2\pi_k - 1)(2\pi_l - 1)| \tag{38}$$

Assuming $(2\pi_i - 1) > \delta > 0$, $\forall i$, the latter simplifies to,

$$s_{ij} = |c^+_{ij}|\delta^2 m^2 (1 - \tfrac{3}{K} + \tfrac{2}{K^2}) > |c^+_{ij}|\delta^2 m^2(1-\tfrac{3}{K}) \tag{39}$$

We next turn to proving an upper bound when $\mathbf{c}(i) \neq \mathbf{c}(j)$. Once again we separate the different elements into five types,

Element type	Number of elements
$\mathbf{c}(i) \neq \mathbf{c}(j) \neq \mathbf{c}(k) \neq \mathbf{c}(l)$	$m^2\left(1-\frac{5}{K}+\frac{6}{K^2}\right)$
$\mathbf{c}(i) \neq \mathbf{c}(j) \neq \mathbf{c}(k) = \mathbf{c}(l)$	$m^2\left(1-\frac{2}{K}\right)\left(\frac{1}{K}-\frac{1}{m}\right)$
$\mathbf{c}(i) \neq \mathbf{c}(j) = \mathbf{c}(k) \neq \mathbf{c}(l)$	$4m^2\left(\frac{1}{K}-\frac{1}{m}\right)\left(1-\frac{2}{K}\right)$
$\mathbf{c}(i) \neq \mathbf{c}(j) = \mathbf{c}(k) = \mathbf{c}(l)$	$2m^2\left(\frac{1}{K}-\frac{2}{m}\right)\left(\frac{1}{K}-\frac{2}{m}\right)$
$\mathbf{c}(k) = \mathbf{c}(i) \neq \mathbf{c}(j) = \mathbf{c}(l)$	$2m^2\left(\frac{1}{K}-\frac{1}{m}\right)\left(\frac{1}{K}-\frac{1}{m}\right)$

The contribution comes from the second,third and fourth types, as according to our model, if all indices come from different groups, or if three come from the same group, the determinant is equal to 0 (see Sec. B). In addition, since $(2\pi_i - 1) > \delta > 0\,\forall i$, the values of r_{ij} are positive for all (i,j) pairs. Since $0 < r_{ij} \leq 1$ for all score elements $s^{kl}_{ij} = |r_{ij}r_{kl} - r_{il}r_{kj}| \leq 1$. The total value of s_{ij} is bounded by the following

$$s_{ij} \leq m^2\left(\frac{1}{K}-\frac{1}{m}\right)\left(5-\frac{8}{K}-\frac{2}{m}\right) < \frac{m^2}{K}\left(5-\frac{8}{K}\right). \tag{40}$$

F Additional results

F.1 Artificial data

In Fig. 7 we present the probability of our spectral clustering based algorithm to recover both the correct number of classes K and the correct assignment function $\mathbf{c}$, as a function of $|G_1|$. Up to $|G_1| = 6$, our algorithm successfully estimates $\mathbf{c}$, with no errors. When $|G_1| > 7$, the performance of the algorithm deteriorates. The degradation in performance presented in Fig. 5, corresponds to the point where the algorithm fails to estimate $\mathbf{c}$ correctly.

In Fig. 8 we present the mean squared error (MSE) in estimating the sensitivities and specificities of the m classifiers, as a function of $|G_1|$, defined as

$$MSE(\{\psi_i, \eta_i\}_{i=1}^m) = \tfrac{1}{2m}\sum_{i=1}^m \left((\hat{\psi}_i - \psi_i)^2 + (\hat{\eta}_i - \eta_i)^2\right). \tag{41}$$

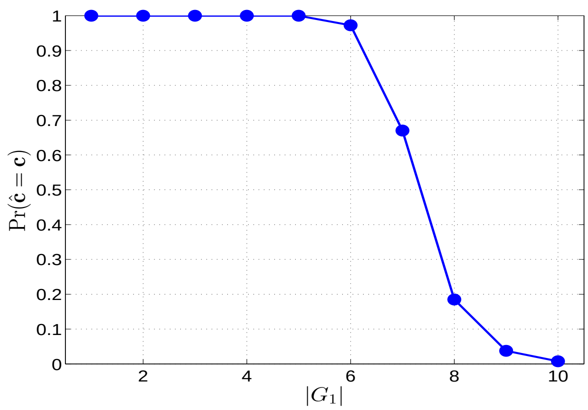

Fig. 7: Probability of estimating the exact assignment function $\mathbf{c}$ as a function of the size of the correlated group $|G_1|$

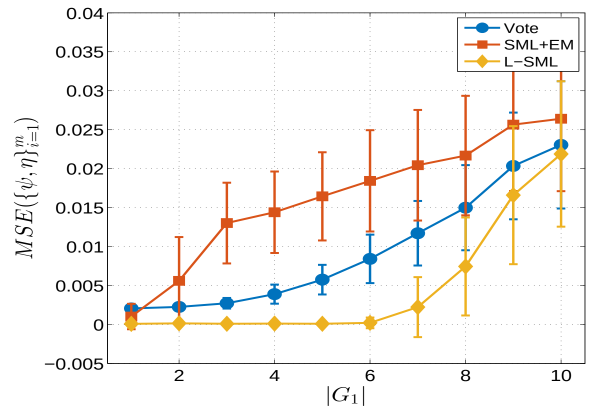

Fig. 8: A comparison of the mean squared error in estimating the accuracies of the different classifiers in the ensemble.

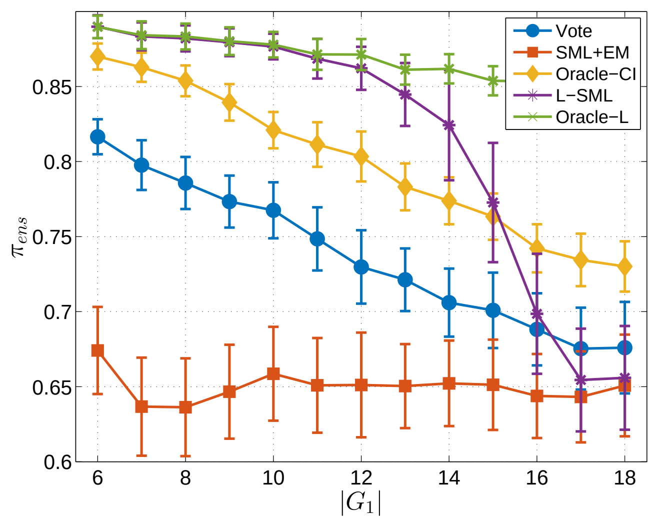

Fig. 9: Artifical data - The total number of classifiers equals $m = 40$. The graph presents the balanced accuracies of several aggregation methods versus the size of the correlated group.

We compare the following three methods: (1) Majority vote ; (2) SML+EM; (3)L-SML. It can be seen that the performance of the SML degrades very fast when the conditional independence assumption is violated. The performance of the L-SML is almost perfect up to the point where $|G_1| = 6$, where as we have seen in Fig. 7, the model is still correctly estimated. The performance is still superior to other methods, even for large values of $|G_1|$. In Fig. 9 we repeat the same experiment described in Sec. 5.1 with $m = 40$ classifiers. Comparing Fig. 9 to Fig. 5 it can be seen that it is not the absolute value of the size of G_1 that determines the performance, but the ratio between $|G_1|$ and the number of classifiers m.

F.2 UCI results

For the magic dataset, Fig. 10 presents the conditional covariance matrix $\frac{1}{2}(C^+ + C^-)$, which is unknown to us. The group of SVM classifiers (12-16) are highly dependent, as well as the group of Naive Base classifiers (8-11). The groups of Random Forest classifiers and logistic model trees are weakly dependent.

Fig. 11 presents an example of the estimated assignment function $\hat{\mathbf{c}}$ for the same dataset. The groups of SVM classifiers were assigned together, as well as the Naive Base classifiers. Except for a single pair, the Random Forest and logistic model trees were assigned to separate groups.

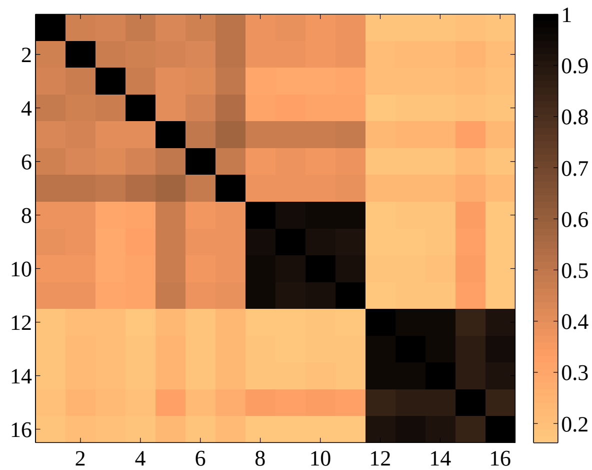

Fig. 10: Magic database - conditional covariance matrix $\frac{1}{2}(C^{+}+C^{-})$.

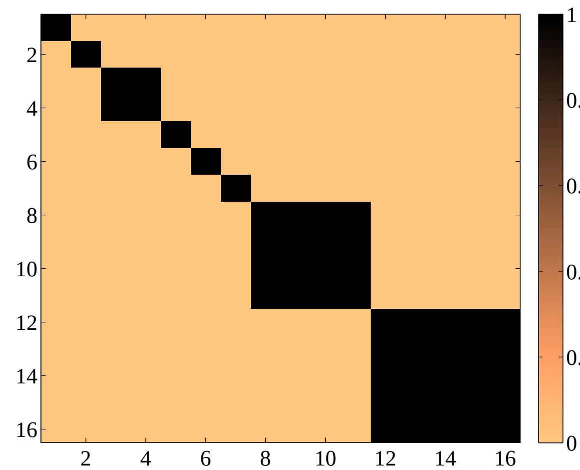

Fig. 11: Magic database - The estimated group return by our algorithm.

In figures 12,13 and 14 we present the results for the following 3 additional datasets from the UCI repository:

- Musk dataset - detection of certain types of molecules.
- Spam dataset - detection of spam from regular mail.
- Miniboo dataset - detection of electron neutrinos (signal) from muon neutrinos (background).

The base classifiers are identical to the ones used for the Magic dataset: (1) 4 Random Forest (2) 3 Logistic Model Trees (3) 3 Naive Bayes (4) 4 SVM .

In figures 12-15, the x-axis is the L-SML balanced error, and the y-axis is the SML balanced error. The results of multiple experiments, each time with the classifiers constructed using different random subset of labeled examples, are presented as blue dots, while the red line represents the $y = x$ line, i.e. when the error of the L-SML and SML are the same. For the Magic dataset, figure 15, we add two lines which represent 2% and 4% improvement over the standard SML.

We can see in the figures that the improvement due to explicit modeling of possible classifier dependencies is consistent across all datasets. The amount of improvement changes, however from dataset to dataset. The following table presents a summary of the different properties of the datasets together with the average improvement in the balanced accuracy between the two methods.

Next, we compare the estimated error rates of the L-SML to the agreement based method (AR) proposed in [19]. Both methods were tested on the magic dataset, with the same ensemble described in Sec. 5.2. Fig. 16 presents the estimated error versus. the true error rate for each of the classifiers in the ensemble. It can be seen that the estimates of both methods contain a bias and are overly optimistic. However, the L-SML successfully identifies the subsets of accurate and inaccurate classifiers.

Dataset	Number of instances	number of features	Mean difference
Magic	19000	11	4%
Spam	4600	57	0.5%
Miniboo	130000	50	0.2%
Musk	6600	168	4.7%

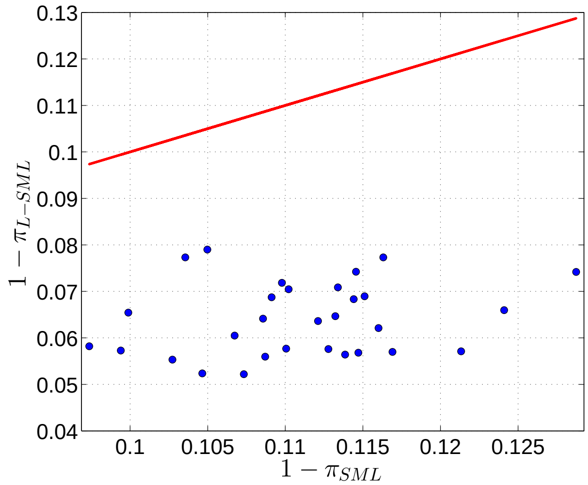

Fig. 12: UCI 'musk' dataset, a comparison between the balanced error of the SML and L-SML.

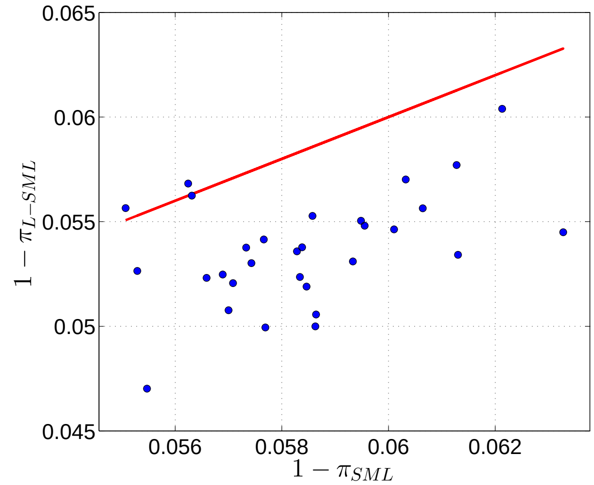

Fig. 13: UCI spambase dataset, a comparison between the balanced error of the SML and L-SML.

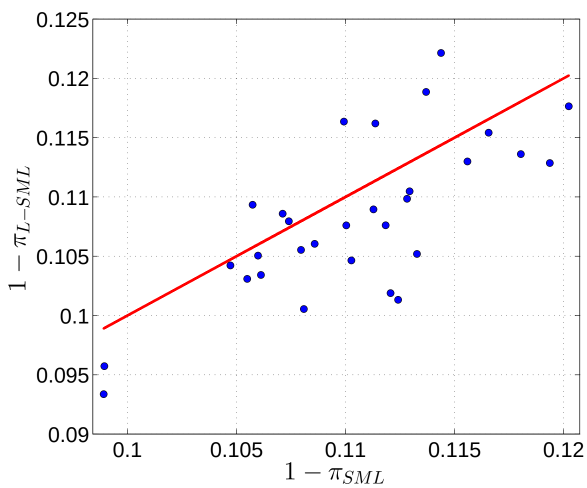

Fig. 14: UCI miniboo dataset, a comparison between the balanced error of the SML and L-SML.

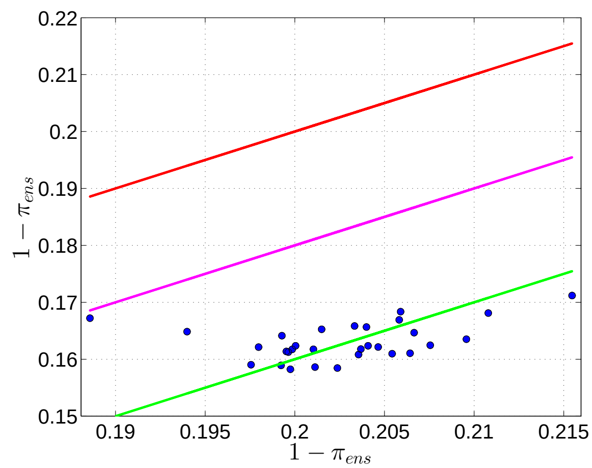

Fig. 15: Magic dataset. The magenta and green lines represent 2% and 4% balanced accuracy improvement over the SML results.

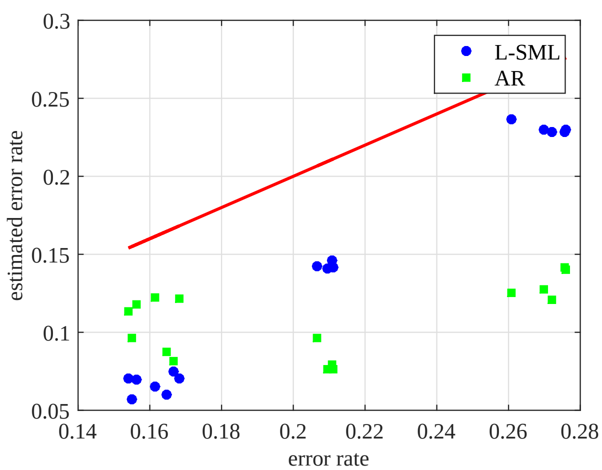

Fig. 16: Comparison to the agreement rates (AR) method presented in [19] on the magic dataset.