Crowdsourcing Regression: A Spectral Approach

Yaniv Tenzer
Weizmann Institute of Science

Omer Dror
Weizmann Institute of Science

Boaz Nadler
Weizmann Institute of Science

Erhan Bilal
IBM T.J. Watson Research Center

Yuval Kluger
Yale University, School of Medicine

Abstract

Merging the predictions of multiple experts is a frequent task. When ground-truth response values are available, this merging is often based on the estimated accuracies of the experts. In various applications, however, the only available information are the experts' predictions on unlabeled test data, which do not allow to directly estimate their accuracies. Moreover, simple merging schemes such as majority voting in classification or the ensemble mean or median in regression, are clearly sub-optimal when some experts are more accurate than others. Focusing on regression tasks, in this work we propose U-PCR, a framework for *unsupervised ensemble regression.* Specifically, we develop spectral-based methods that under mild assumptions and in the absence of ground truth data, are able to estimate the mean squared error of the different experts and combine their predictions to a more accurate meta-learner. We provide theoretical support for U-PCR as well as empirical evidence for the validity of its underlying assumptions. On a variety of regression problems, we illustrate the improved accuracy of U-PCR over various unsupervised merging strategies. Finally, we also illustrate its applicability to unsupervised multi-class ensemble learning.

1 INTRODUCTION

In multiple contemporary applications, there is a need to fuse or merge the predictions of multiple experts or predictors. In this work we consider an unsupervised setting, whereby the experts' accuracies are not a-priori known, and there is no labeled data to estimate it. Instead, the available information are the experts' predictions on unlabeled test data. One notable example of such a setting is crowdsourcing, where the predictions made by multiple human annotators need to be combined (Liu et al., 2019; Rodrigues and Pereira, 2018). Another application domain is biology, where there are extensive collaborative efforts to solve challenging problems by combining the predictions of different research groups. In several past and ongoing DREAM competitions[1], multiple participants construct prediction models based on publicly available labeled data. These are evaluated on held-out data whose statistical distribution may differ significantly from the training one, so participants' accuracies on the training data, even if available, may be misleading. Yet another example is in seismology. Here, the strength of detected earthquakes are estimated at different monitoring stations, based on their distance from the earthquake location and their own measured seismic signals. Next, these station estimates are merged to provide a network estimate of earthquake magnitude. Similar problems also appear in medicine, where there is a need to fuse the results from multiple sources or tests, without a gold standard. In all of these cases, given the experts' predictions on unlabeled test data, key tasks are to estimate their accuracies and provide more accurate predictions than those of the individual experts, by cleverly combining their predictions.

Most prior work in unsupervised ensemble learning considered *discrete* outputs, namely binary, multiclass or ordinal classification (Johnson, 1996; Sheng et al.,

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151. Copyright 2022 by the author(s).

[1] www.dreamchallenges.org

2008; Whitehill et al., 2009; Raykar et al., 2010; Platanios et al., 2014, 2016; Zhou et al., 2012). One of the first works, by Dawid and Skene (1979), assumed that conditioned on the unobserved true class label, experts make independent errors. Despite its simplicity, their model has proven to be very useful in practice. As the likelihood of their model is non-convex, Dawid and Skene (1979) proposed to maximize it by the expectation-maximization algorithm. Recently, Anandkumar et al. (2014); Zhang et al. (2014); Jaffe et al. (2015) proposed computationally efficient and statistically consistent spectral and tensor based methods to address this problem.

In this work we consider an unsupervised ensemble regression setting, involving an explanatory vector X, its real-valued response Y and m experts or predictors f_i. As reviewed in Section 3, most prior work on ensemble regression considered the supervised setting. Only a handful of papers dealt with the unsupervised case, by making quite restrictive modeling assumptions. In contrast, in this work we make much milder assumptions.

As detailed in Section 4, we make the following assumptions, without which unsupervised ensemble regression is in general not possible: (i) the regression problem is learnable, namely it is possible to accurately predict Y given X; (ii) most experts are reasonably accurate and different from each other; and (iii) the mean of Y and a bound on its variance are assumed to be known.

In the unsupervised setting we consider, there is no a-priori knowledge on the mean squared error of the different experts and no labeled data to estimate it. Instead, the available observations are an $m \times n$ matrix of real-valued predictions $f_i(x_j)$ made by the m experts on a set of $n \gg 1$ unlabeled samples $\{x_j\}_{j=1}^n$. Given the matrix $f_i(x_j)$, and the above three assumptions, we aim to (i) detect the most and least accurate experts; and (ii) construct an ensemble predictor for the unobserved responses y_j.

Focusing on linear aggregation methods, in Section 2 we review the optimal weights that minimize the mean squared error (MSE). These depend on two quantities: The $m \times m$ covariance matrix C of the m regressors, and the vector $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_m)^T$ of their covariances with the response. Their entries are given by

$$\begin{aligned} C_{ij} &= \mathbb{E}_X[(f_i(X) - \mu_i)(f_j(X) - \mu_j)] \\ \rho_i &= \mathbb{E}_{(X,Y)}[(f_i(X) - \mu_i)(Y - \theta_1)] \;, \end{aligned} \tag{1}$$

where $\theta_1 = \mathbb{E}[Y]$, and $\mu_i = \mathbb{E}[f_i(X)]$. The matrix C may be estimated from the predictions $f_i(x_j)$ of the m experts. The key challenge is thus to estimate the entries of $\boldsymbol{\rho}$, in the absence of the response values y_j.

Our main contribution, detailed in Section 4, is a spectral framework for unsupervised ensemble regression. Specifically, we develop spectral-based methods that under the above assumptions, are able to estimate the expert accuracies and combine them to an accurate meta-learner. Our approach relies on the following insights: under our modeling assumptions, the covariance matrix C of the m experts can be decomposed as $C = L + S$, where L is low-rank, and S is in general full rank, but is a small perturbation compared to L. Furthermore, up to an unknown constant g_2, the entries of L depend linearly on the vector $\boldsymbol{\rho}$. Hence, our method, denoted U-PCR for Unsupervised Principal Component Regression, consists of four steps, all of which do not require labeled data: (i) extract the low rank matrix L under suitable structural assumptions on S; (ii) estimate the scalar g_2; (iii) given estimates of L and g_2, extract the vector $\boldsymbol{\rho}$; (iv) Given an estimate $\hat{\boldsymbol{\rho}}$, we compute the weights of the linear ensemble learner, also by a spectral-based approach.

To recover the matrix L given C or an estimate thereof, we consider two possible assumptions on the matrix S. Perhaps the simplest assumption is that the deviations of all the m experts from the optimal regressor are pairwise uncorrelated. This implies that $S_{i,j} = 0$ for all $i \neq j$. Under this assumption extracting L is trivial as the off-diagonal entries of C and L coincide. This can be viewed as the regression analogue of the popular Dawid-Skene model in classification.

Nevertheless, the uncorrelated deviations assumption often does not hold in real-life settings. We thus propose a more flexible model, whereby a few experts have correlated deviations. This, in turn, implies that S is *sparse*. Extracting L then boils down to decomposing C into the sum of a low rank plus sparse matrices. By analyzing the specific structure of our problem, we derive sharp conditions on the level of sparsity of S that enables exact recovery of L.

Next we address step (ii) of U-PCR, and derive a model selection procedure to estimate the unknown value g_2. This procedure is justified by a perturbation analysis of the covariance matrix C, assuming that the experts are sufficiently accurate. Our analysis is of independent interest as it also provides a rigorous mathematical support for the supervised ensemble regression method PCR* proposed by Merz and Pazzani (1999). Finally, for step (iii) we show that given estimates of L and g_2, the vector $\boldsymbol{\rho}$ may be estimated by solving a simple system of linear equations.

Section 5 describes experimental results on a variety of problems. This includes both problems for which we trained multiple regression algorithms, as well as real applications where the regressors were constructed by others and only their predictions were available to us. We illustrate that on a variety of real-life regression tasks, our modeling assumptions hold, namely that C can indeed be well approximated by the sum of a low rank and a sparse matrix. Furthermore, as we empirically show, given the predictions $f_i(x_j)$, the mean of Y and a bound on its variance, U-PCR is able to (i) reliably detect the most and least accurate experts; and (ii) predict as accurately as, and often better than, the mean and median of the m regressors. Finally, we illustrate the broader applicability of U-PCR to multi-class unsupervised ensemble learning, provided that the experts output a vector of class probabilities. Section 6 concludes with a summary and discussion.

2 PROBLEM SETUP

Consider a pair of random variables (X, Y), where X belongs to some instance space $\mathcal{X}$ and $Y \in \mathbb{R}$ is its response. Let $\{f_1, \ldots, f_m\}$ be m pre-constructed regression functions, $f_i : \mathcal{X} \to \mathbb{R}$, also called experts or predictors. The task is to construct an ensemble regressor to predict the response y at an instance x by combining the predictions $f_i(x)$ of the m experts.

We consider this ensemble regression problem in an unsupervised setting, whereby the regressors are viewed as black boxes with no knowledge on how they were constructed. In addition, we have no a-priori knowledge on the accuracy or mean squared error of each regressor and no labeled data pairs (x_j, y_j) to estimate it. The available data is an $m \times n$ matrix with the predictions $f_i(x_j)$ of the m experts over n i.i.d. instances $\{x_j\}_{j=1}^n$ from the marginal distribution of X. Given the matrix $f_i(x_j)$, the mean of Y and an upper bound on its variance, but no labeled data, we wish to: (i) estimate the experts' accuracies, and (ii) construct an accurate ensemble learner.

For this unsupervised ensemble regression task to be feasible, beyond knowledge of the mean of Y and a bound on its variance, we also assume that X is informative for the prediction of Y, and that most experts are reasonably accurate and sufficiently diverse. In the next section we make these assumptions more precise and explain how they are utilized in our derivation.

We measure the accuracy of a predictor by its mean squared error, $\text{MSE} = \mathbb{E}[(Y - \hat{y}(X))^2]$. For task (ii) we consider *linear* ensemble learners of the form

$$\hat{y}_{\mathbf{w}}(x) = \theta_1 + \sum_{i=1}^{m} w_i \big(f_i(x) - \mu_i\big). \tag{2}$$

Our goal is to compute a weight vector $\mathbf{w} = (w_1, \ldots, w_m)^T$ so that the corresponding ensemble predictor $\hat{y}_{\mathbf{w}}$ has a small MSE. The following lemma, proven in the supplement, shows that the optimal weights depend only on the covariance C and vector $\boldsymbol{\rho}$, defined in Eq. (1).

Lemma 1. *Any vector $\mathbf{w}^*$, such that $\mathbf{w}^* \in \operatorname{argmin}_{\mathbf{w}} \ \mathbb{E}_{(X,Y)}\big[\big(\hat{y}_{\mathbf{w}}(X) - Y\big)^2\big]$, satisfies*

$$\boldsymbol{\rho} = C\mathbf{w}^*. \tag{3}$$

In particular, if C is invertible then $\mathbf{w}^*$ is unique.

In an unsupervised scenario with $n \gg 1$ samples, the matrix C can be accurately estimated from the predictions $f_i(x_j)$. Similarly, each unknown mean $\mu_i = \mathbb{E}[f_i(X)]$ may be replaced by the empirical mean $\hat{\mu}_i = \frac{1}{n}\sum_j f_i(x_j)$. In contrast, estimating $\boldsymbol{\rho}$ directly by Eq. (1) requires labeled data. The key challenge is thus to find an alternative approach to estimate $\boldsymbol{\rho}$ with no labeled data.

3 PREVIOUS WORK

Most prior work on combining regressors considered supervised settings. For completeness, we briefly review some of these methods, and then discuss unsupervised ensemble approaches.

3.1 Supervised Ensemble Regression

As reviewed in Mendes-Moreira et al. (2012), various supervised ensemble regression approaches were proposed over the past 30 years. Some methods re-train a basic regression algorithm multiple times on different subsets of the labeled data, possibly assigning weights to the labeled instances. Examples include stacking (Wolpert, 1992; Breiman, 1996; Leblanc and Tibshirani, 1996), random forest (Breiman, 2001) and boosting (Freund and Schapire, 1995; Friedman et al., 2000).

Other methods view the regressors as *pre-constructed* and only estimate the weights of their linear combination. In principle, given labeled validation data $\{(x_i, y_i)\}_{i=1}^{n_{\text{val}}}$, one may directly estimate the covariance matrix C and vector $\boldsymbol{\rho}$ of Eq. (1), by their empirical estimators $\hat{C}$ and $\hat{\boldsymbol{\rho}}$. Then, the optimal weights of Eq. (3), may be estimated by $\hat{C}^{-1}\hat{\boldsymbol{\rho}}$. However, due to multi-colinearity of the m experts, the matrix $\hat{C}$ is often ill-conditioned and unstable to invert. To overcome this problem, Merz and Pazzani (1999) proposed a principal component regression approach, called PCR*. In PCR* the weight vector takes the form $\mathbf{w} = \sum_{k=1}^{K} a_k \mathbf{v}_k$, where $\mathbf{v}_k$ are the eigenvectors of $\hat{C}$ and the coefficients a_k are determined by least squares regression over the validation set. The number of components K is chosen by cross validation.

3.2 Unsupervised Ensemble Regression

The simplest unsupervised ensemble methods are the average and the median. By definition, at any instance x the resulting $\hat{y}$ is a function of only $\{f_i(x)\}_{i=1}^m$ and does not depend at all on $f_i(x_j)$ for $x \neq x_j$. Furthermore, these two methods assign equal weights to all experts. Hence, they are in general sub-optimal in heterogeneous situations where some experts are more accurate than others.

Donmez et al. (2010) developed an unsupervised regression method, based on strong parametric assumptions. Specifically, they assumed that the marginal distribution of Y is known and that the m experts follow a known parametric model with parameter θ. Given only unlabeled data, θ can then be estimated by maximum likelihood. Ok et al. (2017) assumed that the response is Gaussian and that each regressor makes a Gaussian error centered around the unknown true response. In contrast, our approach does not assume any parametric model and requires knowing only the mean of Y, and an upper bound on its variance.

Rodrigues and Pereira (2018) proposed to combine the predictions of experts within a deep neural network framework. Specifically, they train a crowd layer neural network using both the original features of each instance x and the predictions of the m experts. Our approach is different (and to some extent simpler) as we only use the predictions $f_i(x)$ and we do not even assume or require access to the original features of each instance x.

More closely related is Wu et al. (2016), who proposed to linearly aggregate the m experts with weights that depend on the leading eigenvector of $\hat{C}$. Their heuristic approach relied on the following assumptions: for any pair $(\mathbf{x}, y)$ there corresponds a binary hidden variable $z \in \{0, 1\}$; each expert f_i follows a two components mixture distribution determined by this hidden variable z; and conditioned on z, the different regressors are independent. In contrast, our approach does not make either of these assumptions. Furthermore, in the simplistic case where experts do make independent errors, our analysis below provides a theoretical support for a variant of their spectral approach.

4 UNSUPERVISED PRINCIPAL COMPONENT REGRESSION

We propose a framework for unsupervised ensemble regression, based on an analysis of the predictions $f_i(x_j)$ and in particular their covariance matrix C. Recall that $\theta_1 = \mathbb{E}[Y]$ is assumed to be known. Hence, without loss of generality we consider mean-centered responses and bias-corrected predictions, $\mathbb{E}[Y] = \mathbb{E}[f_i(X)] = 0$.

We start with our first assumption, that the regression problem is learnable. Let $g(x) = \mathbb{E}[Y|X = x]$ be the conditional mean, which is the optimal predictor of Y given X that minimizes the mean squared error. The predictive ability of X can be quantified by the constant $g_2 = \mathrm{Var}[g(X)] = \mathbb{E}[g(X)Y]$. It is easy to prove that $\mathrm{MSE}[g(x)] = \mathrm{Var}[Y] - g_2$, hence $g_2 \in [0, \mathrm{Var}(Y)]$. A value g_2 close to zero implies that it is not possible to accurately predict Y from X whereas $g_2 = \mathrm{Var}(Y)$ implies perfect error-free prediction. In what follows we assume the problem is learnable in the sense that g_2 is close to $\mathrm{Var}(Y)$. Furthermore, as the value g_2 is typically unknown, we require knowledge of a bound on the variance of Y, so we may estimate g_2 in the interval between zero and this bound, which with some abuse of notation we denote as $\mathrm{Var}[Y]$. We remark that given the assumptions that most experts are reasonably accurate, a bound of $\mathrm{Var}[Y]$ may be estimated from the empirical variances of the m experts.

4.1 Unsupervised PCR

To derive our approach, we write each expert as $f_i(x) = g(x) + h_i(x)$ where $g(x)$ is the conditional mean and h_i is the deviation of f_i from g. Denote $\mathbf{a} \equiv (a_1, \ldots, a_m)$ with entries $a_i = \mathbb{E}[h_i(X)Y]$. Since $\mathbb{E}[g(X)Y] = g_2$, the entries $\rho_i = \mathbb{E}[f_i(X)Y]$ are

$$\rho_i = \mathbb{E}[g(X)Y] + \mathbb{E}[h_i(X)Y] = g_2 + a_i \,. \qquad (4)$$

Similarly, for the entries $C_{i.j} \equiv Cov(f_i(X), f_j(X))$,

$$C_{i,j} = g_2 + a_i + a_j + \mathbb{E}[h_i(X)h_j(X)]. \qquad (5)$$

Equivalently, we may decompose C as

$$C = L + S, \qquad (6)$$

where $L \equiv g_2\mathbf{1}\mathbf{1}^T + \boldsymbol{a}\mathbf{1}^T + \mathbf{1}\boldsymbol{a}^T$ is in general rank two, and $S_{ij} \equiv \mathbb{E}[h_i(X)h_j(X)]$. As we show below, Eq. (6) is crucial for the derivation of U-PCR.

To proceed with the derivation, we assume the experts are sufficiently accurate and different from each other in the following precise sense: the values a_i are much smaller than g_2, which implies that the spectral norm $\|L\| \approx g_2 m$. In addition, we assume that the spectral norm $\|S\| \ll g_2 m$, so the matrix S may be viewed as a small perturbation compared to L.

Our approach then consists of the following four steps: (i) Given the covariance matrix $\hat{C}$, extract an estimate of the low rank matrix L; (ii) Estimate the value of g_2; (iii) Given $\hat{L}$ and $\hat{g}_2$, estimate the vector $\boldsymbol{a}$, which in turn allows to compute $\hat{\boldsymbol{\rho}}$ via Eq. (4); (iv) Given $\hat{C}$ and $\hat{\boldsymbol{\rho}}$, estimate the weight vector $\mathbf{w}$.

Let us first discuss step (iii). The following theorem, proven in the Supplementary, states that if g_2 and the off-diagonal entries of L were known, then this task is indeed feasible.

Theorem 1. *Assume $m \geq 3$. Suppose the off-diagonal entries of L and the value of g_2 are known. Then the vector $\boldsymbol{a}$ is the unique minimizer of*

$$\min_{\tilde{\boldsymbol{a}}} \sum_{i<j} (L_{i,j} - g_2 - \tilde{a}_i - \tilde{a}_j)^2 \tag{7}$$

Importantly, solving (7) with an estimated matrix $\hat{L}$ and an estimated value $\hat{g}_2$, yields a solution $\hat{\boldsymbol{a}}$ and a corresponding estimate $\hat{\boldsymbol{\rho}} = \hat{g}_2 + \hat{\boldsymbol{a}}$ that are stable to errors in L and g_2. Namely, $\|\hat{\boldsymbol{\rho}} - \boldsymbol{\rho}\| = O(\|\hat{L} - L\|, |\hat{g}_2 - g_2|)$. For future use, given an assumed value q for g_2 we denote the resulting solution by $\hat{\boldsymbol{\rho}}(q)$.

We next tackle step (i) of extracting L from the matrix C. We present two possible structural assumptions on S that facilitate this task.

Uncorrelated Errors. The simplest assumption is that the m regressors have *uncorrelated deviations with respect to* $g(X)$,

$$S_{i,j} = \mathbb{E}[h_i(X) h_j(X)] = 0, \quad \forall i \neq j. \tag{8}$$

This is similar to the Dawid-Skene conditional independence model in the classification setting. Under Eq. (8) the off-diagonal entries of C and L coincide.

Beyond Uncorrelated Errors. Assumption (8) is rather restrictive as it requires *all* off-diagonal entries of S to vanish. As this rarely holds in practice, we now present a more flexible model, which allows a few experts to have correlated errors and thus violate Eq. (8). This, in turn, implies that S is sparse.

Let us now describe our approach to extract the vector $\boldsymbol{a}$ under this assumption. To this end, let $vec(S) \equiv (S_{1,2}, S_{1,3}, \ldots, S_{1,m}, S_{2,3} \ldots, S_{m-1,m})^T$ be the *matrix-vectorization* operator that extracts the upper diagonal entries of S, with a similar definition for $vec(C)$. Assuming for the moment that g_2 is known, combining Eq. (6) and the assumption that S is sparse, the vector $\boldsymbol{a}$ may be recovered by solving the following problem,

$$\min_{\mathbf{a}, vec(S)} \{\|vec(S)\|_0 \mid B_1\mathbf{a} + vec(S) = vec(C) - g_2\} \tag{9}$$

where B_1 is the following matrix of size $\frac{m(m-1)}{2} \times m$,

$$B_1 = \left(\begin{array}{c|c|c|c|c|c} 1 & 1 & 0 & \ldots & 0 & 0 \\ 1 & 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 1 & 1 \end{array} \right). \tag{10}$$

Algorithm 1 Sketch of U-PCR

```
Input: Predictions $f_i(x_j), \mathbb{E}[Y]$ and a bound on $\mathrm{Var}(Y)$
Compute $\hat{C}$ and its leading eigenvectors $\mathbf{v}_1, \mathbf{v}_2$
Decompose $\hat{C} = \hat{L} + \hat{S}$
For possible values $q \in [0, \mathrm{Var}(Y)]$ for $g_2$, compute $\hat{\boldsymbol{\rho}}(q)$ by solving Eq. (7)
Compute $\hat{g}_2$ via Eq. (14) and set $\hat{\boldsymbol{\rho}} = \hat{\boldsymbol{\rho}}(\hat{g}_2)$
Compute the weight vector $\mathbf{w}$ via Eq. (15)
Rank the experts by Eq. (16)
For any $x$, output $\hat{y}_{\mathbf{w}}(x)$ by Eq. (2)
```

In general, the solution of (9) is a sparse $vec(S)$ and a dense $\mathbf{a}$, namely it is partially sparse in the optimization variables. Eq. (9) is a generalization of *basis pursuit* whose goal is to find a fully sparse solution (Hastie et al., 2015). Both problems are NP-hard.

A key theoretical question is to quantify how many non-zeros entries can $vec(S)$ have, such that Eq. (9) still admits a unique solution. For the general case with a left hand side of the form $B_1\mathbf{a} + B_2 vec(S)$, this problem was studied by Vaswani and Lu (2010) using the classical notion of restricted isometry property. Here we take advantage of the specific structure of our problem with B_2 being the identity matrix. The following theorem provides a sufficient and necessary condition for the uniqueness of the solution of Eq. (9).

Theorem 2. *Consider a matrix C that follows the decomposition* (6) *with $m \geq 5$ experts. Eq.* (9) *admits a unique solution if and only if $\|vec(S)\|_0 < (m-1)/2$.*

In practice, we only have an estimate of C, and the matrix S is only approximately sparse. Hence, rather than solving Eq. (9), we consider a robust alternative. Concretely, we use the algorithm of Cherapanamjeri et al. (2017) that decomposes an input matrix into a sum of a low-rank and a sparse matrix. Given the resulting matrix $\hat{L}$, we next estimate g_2 as described below, and finally estimate the vectors $\boldsymbol{a}$ and $\boldsymbol{\rho}$, as described above.

Step (ii): Estimating g_2. Extracting the vectors $\boldsymbol{a}$ and $\boldsymbol{\rho}$ assumed that the value $g_2 = \mathrm{Var}(g(X))$ is known. Furthermore, any potential value $q \in [0, \mathrm{Var}(Y)]$ for g_2 leads to a different estimate $\hat{\boldsymbol{\rho}} = \hat{\boldsymbol{\rho}}(q)$. Hence, we now derive a *model selection* criterion to select a value $\hat{g}_2$.

To motivate our proposed estimator of g_2, we utilize our assumption that the experts are quite accurate with S being a small perturbation of L. Concretely, for analysis purposes, it is instructive to scale the deviations h_i by a parameter ϵ,

$$f_i(x) = g(x) + \epsilon h_i(x). \tag{11}$$

Under Eq. (11), the population covariance of the m regressors takes the form

$$C(\epsilon) = g_2 \mathbf{1}\mathbf{1}^T + \epsilon(\boldsymbol{a}\mathbf{1}^T + \mathbf{1}\boldsymbol{a}^T) + \epsilon^2 S \qquad (12)$$

where as before $a_i = \mathbb{E}[h_i(X)Y]$ and $S_{ij} = \mathbb{E}[h_i(X)h_j(X)]$. The next lemma characterizes the leading eigenvalue and eigenvector of C as $\epsilon \to 0$.

Lemma 2. *Let $\lambda_1(\epsilon), \mathbf{v}_1(\epsilon)$ be the largest eigenvalue/eigenvector pair of $C(\epsilon)$. Then, as $\epsilon \to 0$,*

$$\begin{aligned} \lambda_1(\epsilon) &= g_2 m + (2\boldsymbol{a}^T\mathbf{1}) \cdot \epsilon + O(\epsilon^2) \\ \mathbf{v}_1(\epsilon) &= g_2 \mathbf{1} + (\boldsymbol{a} - \tfrac{\boldsymbol{a}^T\mathbf{1}}{m}\mathbf{1}) \cdot \epsilon + O(\epsilon^2) \ . \end{aligned} \qquad (13)$$

The decomposition in Eq. (11) and Eqs. (12)–(13) yield several insights. First, under Eq. (11), $\boldsymbol{\rho} = g_2\mathbf{1} + \epsilon\boldsymbol{a}$. Next, comparing this to Eq. (13), implies that the vector $\boldsymbol{\rho}$ and the leading eigenvector $\mathbf{v}_1$, properly scaled, are *nearly identical*, up to a small shift by $(\frac{1}{m}\sum a_i)\epsilon$ and up to $O(\epsilon^2)$ terms. The following is thus a natural model selection criterion for g_2:

$$\hat{g}_2 = \underset{q \in [0, \mathrm{Var}(Y)]}{\arg\min} \ \mathrm{RES}(q) \equiv \underset{q}{\arg\min} \frac{\|\hat{\boldsymbol{\rho}}(q) - \left(\mathbf{v}_1^T \hat{\boldsymbol{\rho}}(q)\right) \mathbf{v}_1\|}{\|\hat{\boldsymbol{\rho}}(q)\|}. \qquad (14)$$

Step (iv): the weight vector of U-PCR. In principle, given $\hat{\boldsymbol{\rho}}(\hat{g}_2)$ and the estimated covariance matrix $\hat{C}$, by Lemma 1, we could attempt to estimate $\hat{\mathbf{w}}$ via $\hat{C}^{-1}\hat{\boldsymbol{\rho}}$. However, in practice the matrix $\hat{C}$ is often ill-conditioned and thus highly unstable to invert.

To overcome this challenge, note that by our assumption that S is a small pertubation of L, the matrix C is approximately rank two, and spanned by the two vectors $\mathbf{1}$ and $\boldsymbol{a}$. Moreover, up to terms involving the matrix S ($O(\epsilon^2)$ terms in our perturbation analysis), the vector $\boldsymbol{\rho}$ is a linear combination of the first two eigenvectors of C. Therefore, even though the matrix C is ill conditioned, a principal component approach with $K = 2$ components provides an excellent approximation to the optimal weight vector $\mathbf{w}^*$. We hence propose the following weight vector,

$$\mathbf{w}^{\text{U-PCR}} = \frac{1}{\lambda_1}(\mathbf{v}_1^T \hat{\boldsymbol{\rho}}(\hat{g}_2))\, \mathbf{v}_1 + \frac{1}{\lambda_2}(\mathbf{v}_2^T \hat{\boldsymbol{\rho}}(\hat{g}_2))\, \mathbf{v}_2. \qquad (15)$$

A sketch of U-PCR appears in Algorithm 1.

While our focus is on unsupervised ensemble regression, the above analysis is also of independent interest for the supervised case. In particular, it provides theoretical support for the PCR* method of Merz and Pazzani (1999), that also expanded the weight vector as a linear combination of the first few eigenvectors of the predictors' covariance matrix $\hat{C}$.

4.2 Ranking the Experts

In various applications it is of interest to rank the experts by their mean squared errors, and in particular detect the most and least accurate experts. Let us show how this can be done approximately, even in an unsupervised setting. To this end, note that the empirical MSE can be decomposed as follows

$$\begin{aligned} \widehat{\mathrm{MSE}}_i &\equiv \tfrac{1}{n}\sum_j \left(y_j - f_i(x_j)\right)^2 \qquad (16) \\ &= \tfrac{1}{n}\sum_j y_j^2 - \tfrac{2}{n}\sum_j y_j f_i(x_j) + \tfrac{1}{n}\sum_j f_i(x_j)^2. \end{aligned}$$

The first sum on the right hand side is unknown, since the y_j's are unobserved. Yet, it is the same value for all experts f_i. The third sum can be evaluated directly from the observed predictions. Hence, estimating the second sum by $-2\hat{\rho}_i$ allows us to rank the experts.

5 EXPERIMENTS

We illustrate the performance of U-PCR on various real world datasets. These include problems for which we trained multiple regression algorithms as well as applications where the regressors were constructed by a third party and only their predictions were given to us. We denote by IU-PCR and SU-PCR our ensemble learners, based on either the uncorrelated errors or the sparse correlation errors assumptions, respectively.

We compare IU-PCR and SU-PCR to the ensemble mean and median. We also compare to the fully supervised linear oracle regressor of the form (2), which has access to all n response values y_j, computes $\boldsymbol{\rho}_{\text{or}} = \frac{1}{n}\sum y_j f_i(x_j)$, and determines the weight vector by ordinary least squares over the n samples:

$$\mathbf{w}_{\text{or}} = \hat{C}^{-1} \cdot \boldsymbol{\rho}_{\text{or}} \ .$$

5.1 Manually Crafted Ensembles

We considered 17 different regression tasks, including energy output prediction in a power plant, flight delays, basketball scoring and more. Each dataset was randomly split into n_{train} samples used to train $m = 10$ different regression algorithms, including Ridge Regression, SVR, Kernel Regression and Decision Trees, among others. On the remaining n samples, we applied the predictors and constructed the matrix $f_i(x_j)$. For further details, see Table 7 in the Supplementary. We repeated this protocol 20 times for each dataset, each time with a different split of train and test samples.

Following our theoretical analysis, we assess the extent to which the next two conditions hold: (i) the experts' covariance matrix C is approximately low rank; and

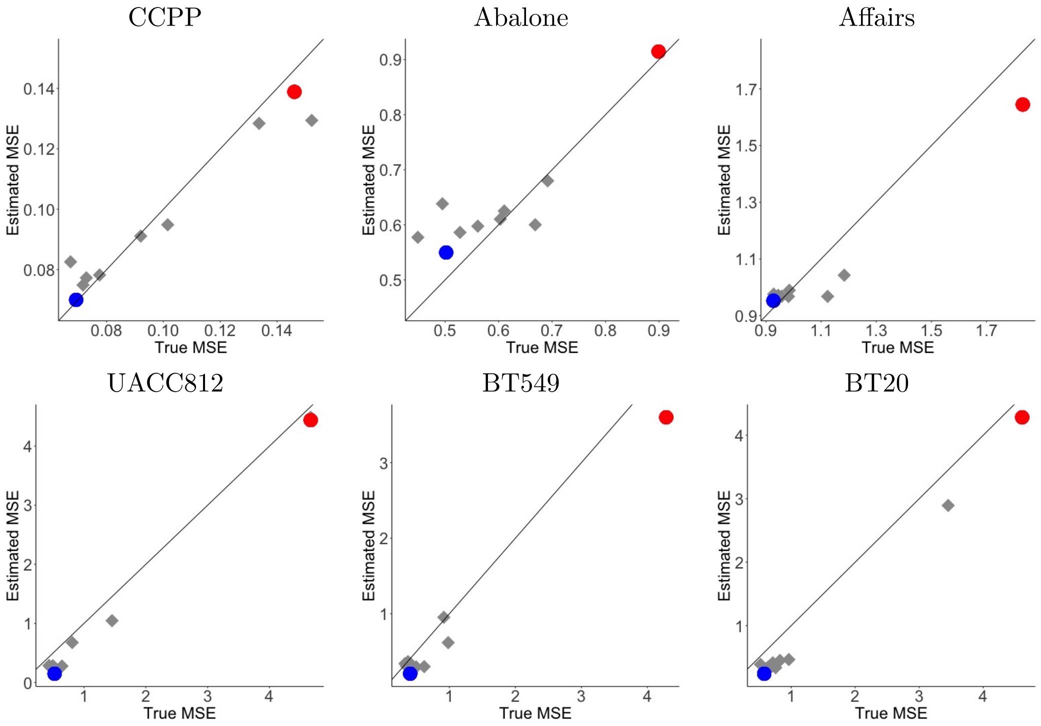

Figure 1: True vs. estimated MSEs normalized by $Var(Y)$, for the CCPP, CO2 and Affairs prediction tasks (top) and for three out of 4 regression tasks in the HPN-DREAM challenge (bottom). The best and worst regressors according to SU-PCR algorithm are marked with blue and red colors respectively.

(ii) the oracle vector $\boldsymbol{\rho}_{\text{or}}$ can be well approximated by its projection on the first two eigenvectors of C. In addition, we present a method to validate our modeling assumption that the matrix S is approximately sparse.

To this end, let λ_1, λ_2 be the first two eigenvalues of $\hat{C}$ and $\mathbf{v}_1, \mathbf{v}_2$ the respective eigenvectors. To quantify the extent to which (i) holds, for each dataset we compute the two ratios $\lambda_1/Tr(\hat{C})$ and $(\lambda_1+\lambda_2)/Tr(\hat{C})$. To assess condition (ii) we project the oracle vector $\boldsymbol{\rho}_{\text{or}}$ onto $\mathbf{v}_1, \mathbf{v}_2$, and report the residual norm divided by $\|\boldsymbol{\rho}_{\text{or}}\|$. Table 1 in the supplement shows that in many datasets both conditions approximately hold. In particular, $\hat{C}$ is indeed nearly low rank with $(\lambda_1+\lambda_2)/Tr(\hat{C}) \in [0.83, 0.99]$. Also,

$$\|\boldsymbol{\rho}_{\text{or}} - (\mathbf{v}_1^T\boldsymbol{\rho}_{\text{or}})\mathbf{v}_1 - (\mathbf{v}_2^T\boldsymbol{\rho}_{\text{or}})\mathbf{v}_2\|/\|\boldsymbol{\rho}_{\text{or}}\| \in [0.01, 0.22].$$

Importantly, in accordance with our theoretical analysis, as seen in Fig. 4, the closer the matrix C is to being rank 2, namely $(\lambda_1+\lambda_2)/Tr(\hat{C}) \approx 1$, the better $\boldsymbol{\rho}_{\text{or}}$ can be approximated by a linear combination of $\mathbf{v}_1, \mathbf{v}_2$. Finally, as can be observed in Tables 1 and 2, for datasets where C was low rank and the above residual was small, SU-PCR tended to be the most accurate ensemble learner.

Next, we consider the assumption of approximate sparsity of S. Since in general the optimal predictor $g(x)$ is unknown, the deviations $h_i(x)$ of each expert f_i from $g(x)$ are also unknown. Hence, directly estimating $S_{i,j} = \mathbb{E}[h_i(X)h_j(X)]$ is impossible. Nonetheless, as we now show, it is possible to assess the validity of this assumption in an unsupervised manner. The idea is to look at the pairwise differences between experts. Define $\Delta_{i,j} \equiv f_i(X) - f_j(X)$, and $\rho_{i,j,k,l} \equiv Cov(\Delta_{i,j}, \Delta_{k,l})$, for $1 \leq i \neq j \neq k \neq l \leq m$. By definition,

$$\begin{aligned}\rho_{i,j,k,l} &= Cov\left((h_i(X) - h_j(X))\,(h_k(X) - h_l(X))\right)\\ &= S_{ik} - S_{il} - S_{jk} + S_{jl}.\end{aligned}$$

Therefore, under the sparsity assumption of S, we expect $\rho_{i,j,k,l}$ to be zero for most quartets of indices. Fig. 3 in the supplement displays the histogram of the empirically estimated $\hat{\rho}_{i,j,k,l}$ for the Abalone, Affair and Flights-JFK datasets. As expected, the resulting histograms are concentrated around zero. Histograms of other datasets were qualitatively similar.

Finally, we evaluate two aspects of the U-PCR algorithm: (i) its model selection criterion for g_2; and (ii) its ability to identify the best and worst regressors via Eq. (16). Due to space limitations, we present only results for SU-PCR. Regarding the model selection criterion, as the true value g_2 is in general unknown, we cannot compute the error $\hat{g}_2 - g_2$. Hence we evaluate the estimated $\hat{g}_2$ by comparing the MSE of the resulting ensemble learner to that attained at other possible values q for g_2. Fig. 2 in the supplement depicts, for three of the datasets, the unobserved MSE(q) obtained by the weight vector of Eq. (15) (blue-solid curve) and the residual RES(q) of Eq. (14) (black-dotted curve), both as a function of the assumed value q for g_2. Also shown are the estimated $\hat{g}_2$ (pink-dashed vertical line), and the value of q that minimizes the MSE curve (red-dashed vertical line). As can be seen, SU-PCR estimated a value $\hat{g}_2$ whose corresponding MSE is close to the minimal MSE achievable by any of the vectors $\hat{\boldsymbol{\rho}}(q)$.

To showcase the ability of U-PCR to detect the best and worst regressors, in the following results we assumed the second moment of Y is known and used Eq. (16) to estimate the MSEs of the various experts. Fig. 1 (top) shows the estimated MSEs vs. the true MSEs. The best and worst regressors are depicted in blue and red, respectively. As can be seen, SU-PCR correctly identified those regressors.

In light of the above results, it might be tempting to choose the best regressor using the estimated MSEs, instead of the U-PCR ensemble learner. Table 2 of the supplement, third and seventh columns (SU-PCR and $\arg\min_i \widehat{\text{MSE}_i}$), show the MSE achieved by SU-PCR and by the single best regressor, respectively, averaged over the 20 train-test random splits of each dataset. As can be seen, in 10 out of 17 prediction tasks, SU-PCR demonstrates favorable performance.

Next we compare IU-PCR and SU-PCR with the ensemble mean and median. As mentioned in Section 3, in the presence of heterogeneity among experts' accuracies, the mean and the median are expected to be sub-optimal, and in general methods that assign different weights to different experts may be more accurate. Indeed, as shown in Table 2, in 12 out of 17 prediction tasks, the MSE attained by SU-PCR is equal or lower than that achieved by the average or the median. In addition, SU-PCR is in general better than IU-PCR attaining lower MSEs in 15 out of the 17 datasets. Figure 5 in the supplement illustrates on three datasets that the estimated MSEs via Eq. (16) of SU-PCR are more accurate than those estimated by IU-PCR. Results on other datasets were qualitatively similar. Finally, the sixth column of Table 2 (Hit-Rate) shows the proportion of realizations at which SU-PCR obtained the minimal MSE. For most datasets and random splits, SU-PCR attains the highest accuracy.

5.2 HPN-DREAM Challenge

The task in the HPN-DREAM breast cancer network inference competition (Hill et al., 2016a) was to predict the time varying concentrations of 4 proteins after introduction of an inhibitor. Understanding the behavior of these proteins is important as it may explain variation in disease phenotypes or therapeutic response (Hill et al., 2016b). Given the predictions of $m = 12$ participants on $n \approx 2500$ instances, we constructed IU-PCR and SU-PCR predictors for each protein. As seen in Fig. 1(bottom), with no labeled data, SU-PCR was able to detect that at least two experts were highly inaccurate. Furthermore, the expert with lowest estimated MSE by SU-PCR is indeed one of the most accurate experts in the ensemble. Due to space limitations, Fig. 1 illustrates this for three of the four prediction tasks. Finally, as shown in Table 3 in the supplement, SU-PCR obtained smaller MSEs than the mean and median on 3 of the 4 proteins, and a comparable MSE on the remaining one.

5.3 *CIFAR10-C* Classification

So far we focused on regression settings. As a last example, we illustrate the broader applicability of the U-PCR framework to multi-class problems. Specifically, we show how one may use the regression technique of this work to perform multiclass unsupervised ensemble learning. The key assumption is that the individual experts output a vector of probabilities, rather than a discrete class label. This is often the case with deep neural networks, whose result is the output of a softmax operation.

Concretely, we consider image classification of the *CIFAR10-C* dataset. This dataset consists of $n = 10^4$ corrupted images from *CIFAR10*, a standard image library with 10 classes. We trained 50 convolutional networks on *CIFAR10* and used the learned models to classify the images of *CIFAR10-C*. As expected, the performance of the various networks sharply decreased when applied to the unseen corrupted images. Indeed, on the test images of *CIFAR10*, the accuracy of the networks was 85% on average, while on the corrupted images of *CIFAR10-C* it dropped to $50\% - 55\%$. This setting is motivated by practical problems in machine learning, where training and test data may have different statistical distributions. Specifically, models are often learned on carefully constructed training data, but then deployed in real world domains, which exhibit different characteristics. Hence, the constructed models may have a much lower (and often unknown) test accuracy compared to their training performance, see D'Amour et al. (2020). Key challenges are to detect which experts are more accurate on the specific test data and to aggregate the original predictors to a more accurate ensemble.

Using the *CIFAR-10C* multiclass problem as an example, we demonstrate how the SU-PCR framework can address these challenges. Here we assume that at any instance x, the i-th expert outputs a vector of probabilities $F_i(x) \in \mathbb{R}^{10}$, whose k-th entry indicates the probability that x belongs to class k. Hence, the data in this case is a tensor Z consisting of 10 separate $m \times n$ matrices. For each class $1 \leq k \leq 10$, the matrix Z^k contains the predicted probabilities for the k-th class, $Z^k_{ij} = F_{i,k}(x_j)$. We adapt the SU-PCR approach to this setting as follows. First, we apply SU-PCR separately to each Z^k, which yields for each instance x_j an estimated probability $p^k(x_j)$ that it belongs to class k. Next, to solve the multi-class problem, we predict the class of an image x_j by $\max_k p^{(k)}(x_j)$. This SU-PCR based approach achieves 62% accuracy in predicting the true class out of the 10 possible class labels. In contrast, majority voting on the predicted labels of the 50 individual networks obtained only 58% accuracy.

6 SUMMARY AND DISCUSSION

We presented a framework to tackle the problem of unsupervised ensemble regression based on the analysis of the experts' covariance matrix. We considered two possible assumptions regarding the statistical dependencies of the deviations of the different experts from the optimal Bayes predictor: (i) an uncorrelated error assumption, which can be viewed as a regression analogue of the Dawid-Skene model in classification; and (ii) a more realistic assumption that the experts' deviations covariance matrix is approximately sparse.

As we demonstrated on a variety of regression tasks, U-PCR was able to detect the best and worst regressors, and construct an ensemble learner more accurate than the mean and median. Finally, we showcased the advantage of the sparsity relaxation over the uncorrelated errors assumption.

Our work raises several directions for future research. One is to extend our approach to a semi-supervised setting, in which there is also a limited amount of labeled data. It is also interesting to theoretically understand the benefits of labeled versus unlabeled data for ensemble learning. Another direction is to explore other relaxations of the uncorrelated error assumption. For example, in unsupervised classification, Fetaya et al. (2016) relaxed the Dawid and Skene conditional independence model by introducing a tree model with an intermediate layer of latent variables. It is interesting whether a similar approach could better model the dependencies in an ensemble of regressors.

Acknowledgements

The authors thank the anonymous referees for constructive suggestions. The authors thank Moshe Guttmann and the Seematics team for their help as well as Pablo Meyer from the Dream Consortium for sharing some of the Dream data. Boaz Nadler is incumbent of the William Petschek Professorial Chair of Mathematics. The work of B.N. was funded by the Intel Collaborative Research Institute for Computational Intelligence, by Pazy Grant ID77-2018 and NIH grant R01GM135928. The work of YK was supported by the National Institutes of Health R01GM131642, UM1PA05141, U54AG076043, R01GM135928, P50CA121974, and U01DA053628

References

Anandkumar, A., Ge, R., Hsu, D., Kakade, S. M., and Telgarsky, M. (2014). Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15(1):2773–2832.

Breiman, L. (1996). Stacked regressions. *Machine Learning*, 24(1):49–64.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Cherapanamjeri, Y., Gupta, K., and Jain, P. (2017). Nearly optimal robust matrix completion. In *International Conference on Machine Learning*, pages 797–805.

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., and Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4):547–553.

D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Chen, C., Deaton, J., Eisenstein, J., Hoffman, M. D., et al. (2020). Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395*.

Dawid, A. P. and Skene, A. M. (1979). Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied statistics*, pages 20–28.

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

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. (2012). The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results.

Fetaya, E., Nadler, B., Jaffe, A., Kluger, Y., and Jiang, T. (2016). Unsupervised ensemble learning with dependent classifiers. In *AISTATS*, pages 351–360.

Freund, Y. and Schapire, R. E. (1995). A decision-theoretic generalization of on-line learning and an application to boosting. In *European conference on computational learning theory*, pages 23–37.

Friedman, J., Hastie, T., and Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting. *The Annals of Statistics*, 28(2):337–407.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical learning with sparsity: the lasso and generalizations*. CRC press.

Hill, S. M., Heiser, L. M., Cokelaer, T., Unger, M., Nesser, N. K., Carlin, D. E., Zhang, Y., Sokolov, A., Paull, E. O., Wong, C. K., et al. (2016a). Inferring causal molecular networks: empirical assessment through a community-based effort. *Nature methods*, 13(4):310–318.

Hill, S. M., Nesser, N. K., Johnson-Camacho, K., Jeffress, M., Johnson, A., Boniface, C., Spencer, S. E., Lu, Y., Heiser, L. M., Lawrence, Y., et al. (2016b). Context specificity in causal signaling networks revealed by phosphoprotein profiling. *Cell systems*.

Jaffe, A., Nadler, B., and Kluger, Y. (2015). Estimating the accuracies of multiple classifiers without labeled data. In *AISTATS*.

Johnson, V. E. (1996). On Bayesian Analysis of Multirater Ordinal Data: An Application to Automated Essay Grading. *Journal of the American Statistical Association*, 91(433):42–51.

Kato, T. (1995). *Perturbation Theory of Linear Operators*. Springer, Berlin, second edition.

Leblanc, M. and Tibshirani, R. (1996). Combining Estimates in Regression and Classification. *Journal of the American Statistical Association*, 91(436):1641–1650.

Lichman, M. (2013). UCI machine learning repository.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755.

Liu, T., Venkatachalam, A., Sanjay Bongale, P., and Homan, C. (2019). Learning to predict population-level label distributions. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 1111–1120.

Mendes-Moreira, J., Soares, C., Jorge, A. M., and Sousa, J. F. D. (2012). Ensemble approaches for regression: A survey. *ACM Computing Surveys (CSUR)*, 45(1):10.

Merz, C. and Pazzani, M. (1999). A Principal Components Approach to Combining Regression Estimates. *Machine Learning*, 36(1-2):9–32.

Nadler, B. (2008). Finite sample approximation results for principal component analysis: A matrix perturbation approach. *The Annals of Statistics*, pages 2791–2817.

Ok, J., Oh, S., Shin, J., Jang, Y., and Yi, Y. (2017). Efficient learning for crowdsourced regression. *arXiv preprint arXiv:1702.08840*.

Platanios, E. A., Blum, A., and Mitchell, T. (2014). Estimating accuracy from unlabeled data. In *Uncertainty in Artificial Intelligence*, pages 682–691.

Platanios, E. A., Dubey, A., and Mitchell, T. (2016). Estimating accuracy from unlabeled data: A bayesian approach. In *International Conference on Machine Learning*, pages 1416–1425.

Raykar, V. C., Yu, S., Zhao, L. H., Valadez, G. H., Florin, C., Bogoni, L., and Moy, L. (2010). Learning from crowds. *Journal of Machine Learning Research*, 11(Apr):1297–1322.

Rodrigues, F. and Pereira, F. (2018). Deep learning from crowds. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Sheng, V. S., Provost, F., and Ipeirotis, P. G. (2008). Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 614–622. ACM.

Vaswani, N. and Lu, W. (2010). Modified-cs: Modifying compressive sensing for problems with partially known support. *IEEE Transactions on Signal Processing*, 58(9):4595–4607.

Whitehill, J., Wu, T.-f., Bergsma, J., Movellan, J. R., and Ruvolo, P. L. (2009). Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in neural information processing systems*, pages 2035–2043.

Wolpert, D. H. (1992). Stacked generalization. *Neural networks*, 5(2):241–259.

Wu, D., Lawhern, V. J., Gordon, S., Lance, B. J., and Lin, C.-T. (2016). Spectral meta-learner for regression (smlr) model aggregation: Towards calibrationless brain-computer interface. In *IEEE International Conference on Systems, Man, and Cybernetics*, pages 743–749.

Zhang, Y., Chen, X., Zhou, D., and Jordan, M. I. (2014). Spectral methods meet em: A provably optimal algorithm for crowdsourcing. In *Advances in neural information processing systems*, pages 1260–1268.

Zhou, D., Basu, S., Mao, Y., and Platt, J. C. (2012). Learning from the wisdom of crowds by minimax entropy. In *Advances in Neural Information Processing Systems*, pages 2195–2203.

A Proofs

Proof of Theorem 1. Following the notations defined in the main text body, the off-diagonal entries of the matrix L have the following structure:

$$L_{ij} = g_2 + a_i + a_j. \tag{17}$$

Since L is symmetric and g_2 is assumed to be known, the off-diagonal entries provide $\binom{m-1}{2}$ linear equations for the m unknown variables $\tilde{\boldsymbol{a}} = (\tilde{a}_1, \ldots, \tilde{a}_m)$. Define the following set of pairwise indices $\mathcal{I} \equiv \{(i, i+1) : 1 \leq i \leq m-1\} \cup \{(i', j')\}$, for some $j' \neq i'+1$. It is easy to prove that if $m \geq 3$, the following is a set of m linearly independent equations, with m variables:

$$L_{ij} - g_2 = \tilde{a}_i + \tilde{a}_j,\, (i,j) \in \mathcal{I} \tag{18}$$

The system above has a unique solution, as it is of a full rank, and by construction it equals to $\boldsymbol{a}$. Next, define $\psi(\tilde{\boldsymbol{a}}) \equiv \sum_{(i,j)\in\mathcal{I}} (L_{i,j} - g_2 - \tilde{a}_i - \tilde{a}_j)^2$ and consider the corresponding quadratic optimization problem

$$\min_{\tilde{\boldsymbol{a}}} \psi(\tilde{\boldsymbol{a}}).$$

Since this function is non-negative, by the arguments above, its minimum is attained at the true vector $\boldsymbol{a}$ for which $\psi(\boldsymbol{a}) = 0$. We next show that $\boldsymbol{a}$ is also the unique minimizer of the following optimization problem

$$\min_{\tilde{\boldsymbol{a}}} \sum_{(i,j):i\neq j} (L_{i,j} - g_2 - \tilde{a}_i - \tilde{a}_j)^2. \tag{19}$$

Assume by contradiction that there exists another solution $\mathbf{a}'$ for the problem above, with a zero objective value. Since for any vector $\tilde{\boldsymbol{a}}$,

$$\sum_{(i,j):i\neq j} (L_{i,j} - g_2 - \tilde{a}_i - \tilde{a}_j)^2 \geq \sum_{(i,j)\in\mathcal{I}} (L_{i,j} - g_2 - \tilde{a}_i - \tilde{a}_j)^2 \tag{20}$$

the vector $\mathbf{a}'$ must be also a solution for the linear system of (18). However, this yields a contradiction to the uniqueness of $\boldsymbol{a}$. The vector $\boldsymbol{\rho}$ can then be computed from Eq. (4), as g_2 is assumed to be known.

□

Proof of Lemma 1. For any estimator $\hat{y}(X)$, opening the brackets in the definition of its MSE gives

$$\mathbb{E}_{(X,Y)}[(Y - \hat{y}(X))^2] = \theta_2 - 2\mathbb{E}_{(X,Y)}[Y \cdot \hat{y}(X)] + \mathbb{E}_X[(\hat{y}(X))^2]$$

where $\theta_2 = \mathbb{E}_Y[Y^2]$. For linear estimators as in Eq. (2), $\hat{y}(X) = \hat{y}_{\mathbf{w}}(X)$,

$$\begin{aligned}
\mathrm{MSE}(\mathbf{w}) &= \theta_2 - 2\mathbb{E}\Big[Y\big(\theta_1 + \sum_i w_i(f_i(X) - \mu_i)\big)\Big] + \mathbb{E}\Big[\big(\theta_1 + \sum_i w_i(f_i(X) - \mu_i)\big)^2\Big] \\
&= \theta_2 - 2\theta_1^2 - 2\sum_i w_i\rho_i + \theta_1^2 + 2\theta_1 \sum_i w_i \mathbb{E}[f_i(X) - \mu_i] + \sum_{ij} w_i w_j C_{ij} \\
&= \mathrm{Var}(Y) - 2\mathbf{w}^T\boldsymbol{\rho} + \mathbf{w}^T C\mathbf{w}.
\end{aligned} \tag{21}$$

The minimal MSE achievable by a linear ensemble regressor, denoted MSE*, is found by minimizing Eq. (21) with respect to the weights $\mathbf{w}$. This gives

$$\boldsymbol{\rho} = C\mathbf{w}^*, \quad \text{and} \quad \mathrm{MSE}^* = \mathrm{Var}(Y) - (\mathbf{w}^*)^T C\mathbf{w}\ .$$

□

To prove Theorem 2 we shall use the following auxiliary lemma.

Lemma 3. *Assume $m \geq 5$. Any non-trivial linear combination of the m columns of the matrix B_1 given in Eq.* (10)*, yields a vector with at least $m-1$ non-zero entries. Namely, $\forall \mathbf{a} \in \mathbb{R}^m$ with $\mathbf{a} \neq \mathbf{0}$, it follows that $\|B_1\mathbf{a}\|_0 \geq m-1$.*

Proof of Lemma 3. For any non-zero vector $\mathbf{a} \in \mathbb{R}^m$, denote by $\mathbf{v} = B_1\mathbf{a}$ and by $q(\mathbf{a})$ the number of positive entries in $\mathbf{a}$. Due to the special structure of B_1, any two of its columns intersect at exactly one entry, and the intersection of any subset of three or more columns is empty. In other words, each entry in $\mathbf{v}$, say k, is the sum of exactly two terms: $v_k = a_i B_1(i,k) + a_j B_1(j,k)$, for some i, j. Since all entries $B(i,j) \in \{0,1\}$, for v_k to vanish, a_i and a_j must have opposite signs. Hence, the total number of zero entries in $\mathbf{v}$ is upper bounded by the number of pairs of indices of $\mathbf{a}$ with opposite signs, namely $q(\mathbf{a})(m - q(\mathbf{a}))$. Define $\psi(q) \equiv q(m-q)$. For m even this function attains its maximum at $q = m/2$ where $\psi(\frac{m}{2}) = m^2/4$. For m odd, the maximum value of $(m+1)(m-1)/4$ is attained at $q = (m \pm 1)/2$. The number of non-zero entries, $\|B_1\mathbf{a}\|_0$ is thus lower bounded by $m(m-1)/2 - m^2/4$ for m even and a similar expression for m odd. It can be easily verified that when $m \geq 5$, these lower bounds are greater or equal to $m - 1$. □

Note that Lemma 3 is tight. Let $\mathbf{u} \in \mathbb{R}^m$ be the vector with alternating ±1 signs, $u_i = (-1)^i$. Then $\|B_1\mathbf{u}\|_0 = m - 1$.

Proof of Theorem 2. Let $\mathbf{a}, vec(S)$ be the true values in the decomposition of C with $\|vec(S)\|_0 < (m-1)/2$. Assume to the contrary that there exists another solution $\tilde{\mathbf{a}}, vec(\tilde{S})$ to Eq. (9), with $\|vec(\tilde{S})\|_0 \leq \|vec(S)\|_0$. Then

$$B_1(\mathbf{a} - \tilde{\mathbf{a}}) = vec(\tilde{S}) - vec(S).$$

Since $\tilde{\boldsymbol{a}} \neq \boldsymbol{a}$, if follows from Lemma 3 that $\|B_1(\mathbf{a} - \tilde{\mathbf{a}})\|_0 \geq m - 1$. However, $\|vec(\tilde{S}) - vec(S)\|_0 \leq 2\|vec(S)\|_0 < m - 1$, a contradiction. For the other direction, we show that there exists a pair $\mathbf{a}, vec(S)$ with $\|vec(S)\|_0 \geq (m-1)/2$ for which Eq. (9) does not have a unique solution. This stems from the fact that Lemma 3 is tight. Specifically, for any such $vec(S)$ choose any $(m-1)/2$ non-zero entries of $B_1\mathbf{u}$ where $\mathbf{u}$ is the vector of alternating ±1 signs, and let $vec(\tilde{S})$ be the remaining non zero entries in this vector. □

Proof of Lemma 2. The proof follows a perturbation approach similar to the one outlined in Nadler (2008). Since $C(\epsilon)$ is symmetric and quadratic in ϵ, classical results on perturbation theory (Kato, 1995) imply that in a small neighborhood of $\epsilon = 0$, the leading eigenvalue and eigenvector are analytic in ϵ. We may thus expand them in a Taylor series,

$$\begin{aligned} \lambda(\epsilon) &= \lambda_0 + \lambda_1\epsilon + \lambda_2\epsilon^2 + \ldots \\ \mathbf{v}(\epsilon) &= \mathbf{v}_0 + \mathbf{v}_1\epsilon + \mathbf{v}_2\epsilon^2 + \ldots \end{aligned}$$

We insert this expansion into the eigenvector equation $C(\epsilon)\mathbf{v}(\epsilon) = \lambda(\epsilon)\mathbf{v}(\epsilon)$ and solve the resulting equations at increasing powers of ϵ.

The leading order equation reads $g_2\mathbf{1}\mathbf{1}^T\mathbf{v}_0 = \lambda_0\mathbf{v}_0$, which gives $\mathbf{v}_0 \propto \mathbf{1}$ and $\lambda_0 = g_2\|\mathbf{1}\|^2 = g_2 m$. Since the eigenvector $\mathbf{v}(\epsilon)$ is defined only up to a multiplicative factor, we conveniently chose it to be such that $\mathbf{1}^T\mathbf{v}(\epsilon) = g_2 m$ holds for all ϵ. This gives $\mathbf{v}_0 = g_2\mathbf{1}$ and $\mathbf{v}_1^T\mathbf{v}_0 = 0$.

The $O(\epsilon)$ equation reads

$$g_2\mathbf{1}\mathbf{1}^T\mathbf{v}_1 + (\boldsymbol{a}\mathbf{1}^T + \mathbf{1}\boldsymbol{a}^T)\mathbf{v}_0 = \lambda_0\mathbf{v}_1 + \lambda_1\mathbf{v}_0. \tag{22}$$

Multiplying this equation from the left by $\mathbf{v}_0^T$ gives

$$2(\mathbf{v}_0^T\mathbf{1})(\boldsymbol{a}^T\mathbf{v}_0) = \lambda_1\|\mathbf{v}_0\|^2$$

or $\lambda_1 = 2\sum a_j$. Inserting the expression for λ_1 back into Eq. (22) gives

$$\mathbf{v}_1 = \frac{1}{\lambda_0}[(\boldsymbol{a}^T\mathbf{v}_0)\mathbf{1} + (\mathbf{1}^T\mathbf{v}_0)\boldsymbol{a} - (2\sum_j a_j)\mathbf{v}_0]$$

from which Eq. (13) readily follows. □

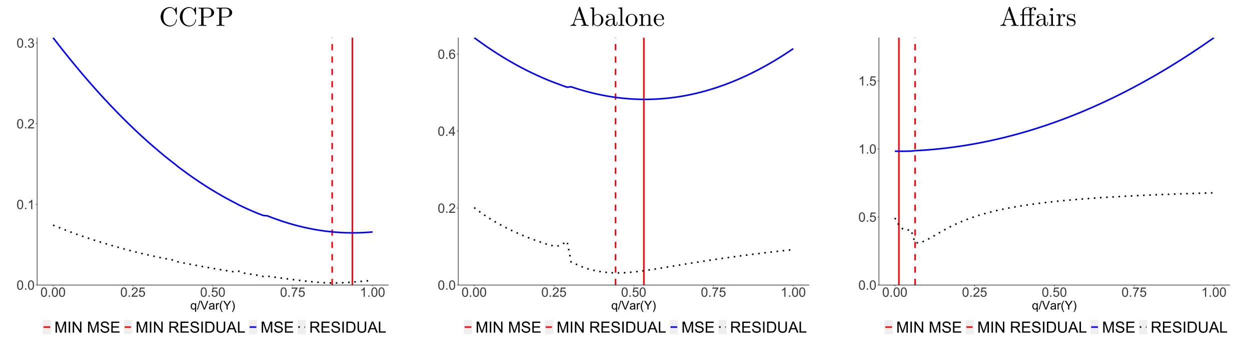

Figure 2: Plots of normalized values of MSE(q) and RES(q) for three different datasets. The solid and dashed vertical lines correspond to the value of $\hat{g}_2$ that minimizes the MSE and residual curves respectively.

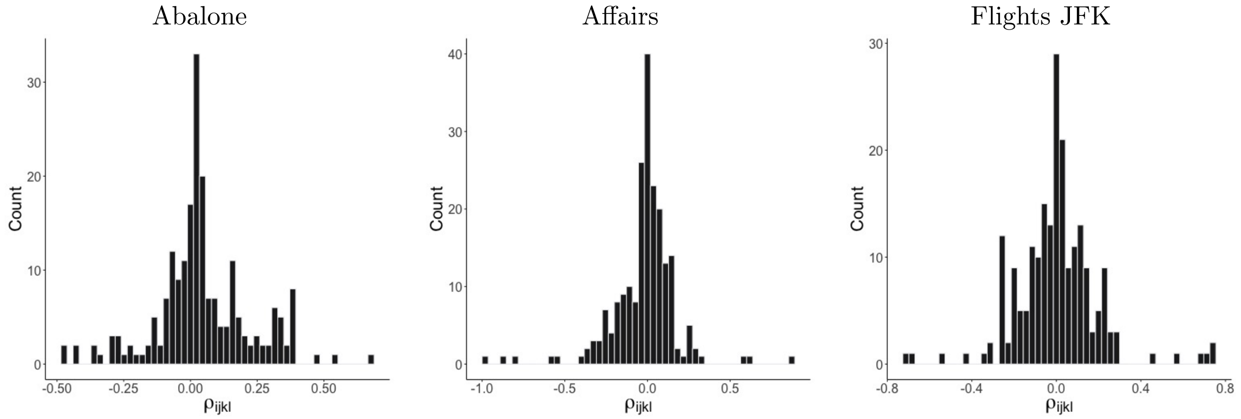

Figure 3: Histogram of $\rho_{i,j,k,l} \equiv Cov\left(\left(h_i(X) - h_j(X)\right)\left(h_k(X) - h_l(X)\right)\right)$ for the Abalone, Affairs and Flights JFK datasets.

B Beyond Uncorrelated Errors - Implementation Details

As mentioned in the main text, we go beyond the simple case of uncorrelated errors by using the projected gradient algorithm of Cherapanamjeri et al. (2017). In a nutshell, the algorithm decomposes a partially observed and corrupted matrix M, into the sum of low rank and sparse matrices, L and S respectively. Our setting can be viewed as a specific instance of the general setting considered in Cherapanamjeri et al. (2017), where the set of observed indices consists of all off-diagonal entries in the empirical covariance matrix $\hat{C}$.

Concretely the algorithm has two main parameters: (i) incoherence parameter μ; (ii) threshold η_t. In all experiments we approximated μ by $\max_{1 \leq i \leq m} \sqrt{\frac{m}{2}} ||e_i^T V||_2$, where V is the matrix whose columns are the eigenvectors of $\hat{C}$. As for η, we follow the authors' recommendation and at each iteration t, we set $\eta = \mu ||M - S_t||_2 / n$, where S_t is the approximated sparse component at iteration t.

C Experiments

Manually Crafted Ensembles - Results

Fig. 2 depicts the unobserved MSE(q) obtained by the weight vector of Eq. (15) in the main text, as a function of the assumed value q for g_2 (solid blue curve) and the residual RES(q) given by Eq. (14) in the main text body (dotted black curve), as a function of $q/\text{Var}(Y)$. In addition shown are the estimated $\hat{g}_2$ (pink-dashed vertical line), and the value of q that minimizes the MSE curve (red-dashed vertical line).

The five leftmost columns of Table 2 show, for each dataset, the MSE achieved by the oracle and each aggregation method, normalized by Var(Y), averaged over the 20 train-test splits. The seventh column of Table 2 presents

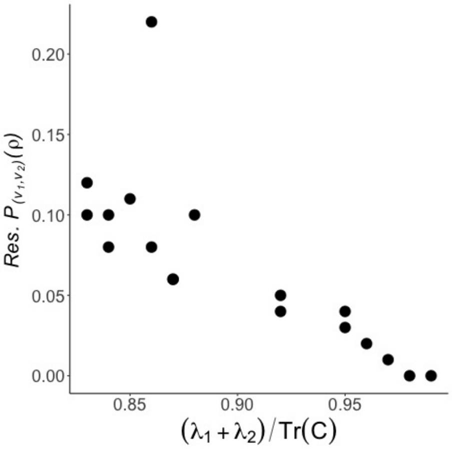

Figure 4: Normalised residual norm $\mathcal{P}_{(v_1,v_2)}(\boldsymbol{\rho})$ vs $(\lambda_1+\lambda_2)/Tr(\hat{C})$, of the various manually crafted ensembles for the 17 datasets.

the MSE achieved by the single best regressor, averaged over the 20 train-test random splits of each dataset. Finally, the sixth column of Table 2 (Hit-Rate in the table) shows the proportion of repetitions at which SU-PCR obtained the minimal MSE.

Figure 5 shows the true MSEs vs. the MSEs estimated by the IU-PCR and SU-PCR (first and second rows, respectively), for the CPP, Wine quality white and abalone datasets.

The HPN-DREAM Challenge Experiment - Results

As described in the main text, the task in the HPN-DREAM breast cancer network inference competition (Hill et al., 2016a) was to predict the time varying concentrations of 4 proteins after introduction of an inhibitor. Given the predictions of $m = 12$ participants on $n \approx 2500$ instances, we constructed separate IU-PCR and SU-PCR predictors for each protein. Table 3 shows the MSEs achieved by the various aggregation methods, normalized by $\mathrm{Var}(Y)$.

Bounding Box Challenge

We present the results on an additional regression task, not discussed in the main text. Here we were given predictions of 16 deep learning models trained by Seematics Inc., on the location of physical objects in images. The models were trained on the PASCAL Visual Object Classes dataset (Everingham et al., 2012), whereas the predictions were made on images from COCO dataset (Lin et al., 2014). For demonstration we focused on three object classes, {Person, Dog, Cat}, with each neural network providing four coordinates for the bounding box: (x_1, y_1) and (x_2, y_2). We consider predicting each of the 4 coordinates as a separate task. Hence, there are a total of 12 regression problems. For each such problem we constructed IU-PCR and SU-PCR ensemble predictors, with the mean squared error as our measure of accuracy. Fig. 6 depicts the true MSE vs. the MSE as estimated by SU-PCR, for coordinate x_1. Results for the other coordinates were qualitatively similar. As shown, in the case of the Cat and Dog classes, the ability of SU-PCR to identify accurate and inaccurate experts is impressive. However, In the case of the Person class, while it succeeds in identifying the worst regressor, the predictor that SU-PCR marks as best, is ranked in the seventh place, according to the true MSE. Checking the rank of matrix $\hat{C}$ reveals that in the case of the Person class, $(\lambda_1+\lambda_2)/Tr(\hat{C}) = 0.81$, while $(\lambda_1+\lambda_2)/Tr(\hat{C}) = 0.93$ for the Cat class and $(\lambda_1+\lambda_2)/Tr(\hat{C}) = 0.97$ for the Dog class. Similarly, n the case of the Person class, $\mathcal{P}_{(v_1,v_2)}(\boldsymbol{\rho}) = 0.093$, while $\mathcal{P}_{(v_1,v_2)}(\boldsymbol{\rho}) = 0.012$ for the Cat class and $\mathcal{P}_{(v_1,v_2)}(\boldsymbol{\rho}) = 0.014$ for the Dog class. Therefore it is apparent that in the Person class, the covariance matrix C demonstrates a stronger deviation from the Low-rank plus Sparse decomposition, than in the other classes. The accuracy of the three object classes and all coordinates are shown in Tables 4, 5 and 4. Note that all values are normalized by $\mathrm{Var}(Y)$. As can be seen both IU-PCR and SU-PCR attain similar or lower MSE compared to the mean and median aggregations in 10 out of 12 datasets. In all prediction tasks, SU-PCR achieves MSEs that are similar or lower than IU-PCR.

Table 1: Empirical validation of the two key conclusions of our theoretical analysis: (i) the matrix C is approximately rank 2 and (ii) $\boldsymbol{\rho}$ can be well approximated by its projection on the the first two eigenvectors of C with a small residual. The affairs dataset is an outlier, for which we indeed were not able to construct an accurate linear ensemble learner, see right panel of Fig. 2.

Dataset	$\lambda_1/Tr(\hat{C})$	$(\lambda_1+\lambda_2)/Tr(\hat{C})$	Res. $\mathcal{P}_{v_1}(\boldsymbol{\rho})$	Res. $\mathcal{P}_{(v_1,v_2)}(\boldsymbol{\rho})$
abalone	0.82	0.91	0.06	0.04
affairs	0.70	0.86	0.57	0.22
basketball	0.89	0.94	0.04	0.04
blog feedback	0.79	0.91	0.06	0.05
CCPP	0.97	0.98	0.04	0.04
CO2	0.95	0.99	0.11	0.04
crime-1	0.90	0.96	0.02	0.01
crime-2	0.84	0.94	0.05	0.03
flights AUS	0.71	0.87	0.10	0.06
flights BOS	0.69	0.85	0.19	0.08
flights BWI	0.68	0.84	0.15	0.10
flights HOU	0.68	0.85	0.26	0.11
flights JFK	0.68	0.84	0.19	0.13
flights LGA	0.66	0.83	0.20	0.12
flights longhaul	0.76	0.88	0.24	0.10
online videos	0.87	0.96	0.07	0.02
qsar aquatic toxicity	0.70	0.84	0.12	0.08
wine quality white	0.72	0.87	0.10	0.06

Manually Crafted Ensembles - Details

For reproducibility, we provide below a short description and reference for every prediction task used in Section 5.1. Table 7 summarizes the main characteristics of each dataset, including the number of randomly chosen samples for training and testing.

Abalone. A dataset containing features of abalone, where the goal is to predict its age (Lichman, 2013). Source: archive.ics.uci.edu/ml/datasets/Abalone

Affairs. A dataset containing features describing an individual such as time at work, time spent with spouse, and time spent with a paramour. The goal here is to predict the time spent in extramarital affairs. Source: statsmodels.sourceforge.net/0.6.0/datasets/generated/fair.html

Basketball. Dataset contains stats on NBA players. Task: Predict number of points scored by the player on the next game. The features are: `name`, `venue`, `team`, `date`, `start`, `pts_ma`, `min_ma`, `pts_ma_1`, `min_ma_1, pts`, where `start` is whether or not the player started, `pts` is number of points scored, `min` is number of minutes played, `ma` stands for moving average, starts at season, and `ma_1` is a moving average with a 1 game lag. Source: http://www.quantifan.com/

Blog Feedback. Instances in this dataset contain features extracted from blog posts. The task associated with the data is to predict how many comments the post will receive. Source: https://archive.ics.uci.edu/ml/datasets/BlogFeedback

Flights. Information on flights from 2008, where the task is to predict the delay upon arrival in minutes. The features here are the date, day of the week, scheduled and actual departure times, scheduled arrival times, flight ID, tail number, origin, destination, and distance. Due to its size, we split this dataset to flights originating from specific airports (AUS, BOS, BWI, HOU, JFK, and LGA), and long-haul flights. Source: statcomputing.org/dataexpo/2009/the-data.html

CCPP. Combined Cycle Power Plant UCI-dataset containing physical characteristics such as temperature and humidity. The task here is to predict the net hourly electrical energy output of the plant. Source: archive.ics.uci.edu/ml/datasets/Combined+Cycle+Power+Plant

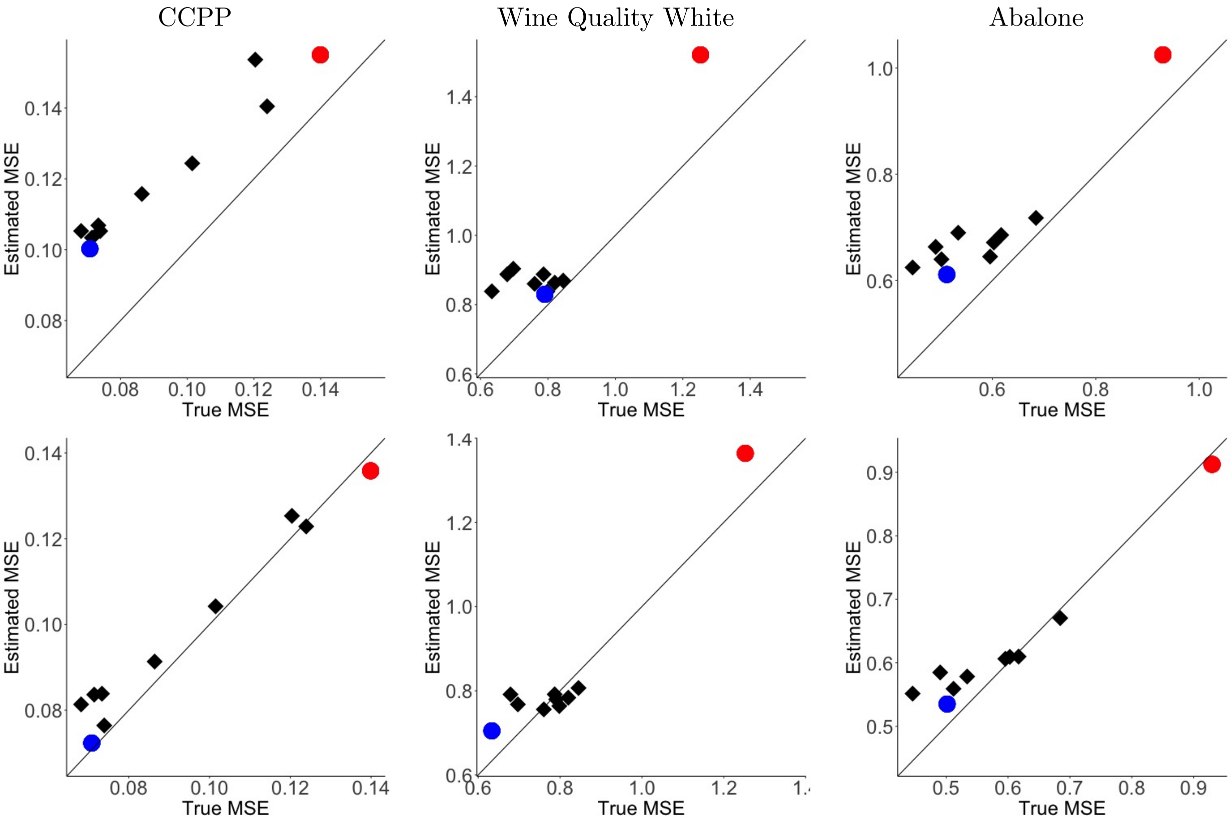

Figure 5: (Top): True MSEs vs. MSEs estimated by IU-PCR (up) and SU-PCR (bottom), all normalized by $Var(Y)$, for the CCPP, Wine quality white and Abalone datasets. The best and worst regressors according to the estimated MSEs are marked with green and red colors respectively.

Online Videos. YouTube video transcoding dataset. Predict the transcoding time based on parameters of the video. Source: archive.ics.uci.edu/ml/datasets/Online+Video+ Characteristics+and+Transcoding+Time+Dataset

Wine Quality White. Predict the quality score (1-10) of white wine based on chemical characteristics, such as acidity and pH level (Cortez et al., 2009). Source: archive.ics.uci.edu/ml/datasets/Wine+Quality

Crime-1. Predict the number of murders per 100K population. Explanatory variables include social characteristics, such as the percent of the population considered urban and the median family income, as well as law enforcement, such as per capita number of police officer and percent of officers assigned to drug units. Source: archive.ics.uci.edu/ml/datasets/Communities+and+Crime+Unnormalized

Crime-2. Predict the number of rapes per 100K population. Explanatory variables include social characteristics, such as the percent of the population considered urban and the median family income, as well as law enforcement, such as per capita number of police officers, and percent of officers assigned to drug units. Source: archive.ics.uci.edu/ml/datasets/Communities+and+Crime+Unnormalized

CO2 Data from an experiment on the cold tolerance of the grass species *Echinochloa crus-galli*. Predict the carbon dioxide uptake rates. Source: the *nmle* package, *R*.

Qsar Aquatic Toxicity Predict the acute aquatic toxicity towards the fish Pimephales promelas, based on 8 molecular descriptors. Source: archive.ics.uci.edu/ml/datasets/QSAR+aquatic+toxicity

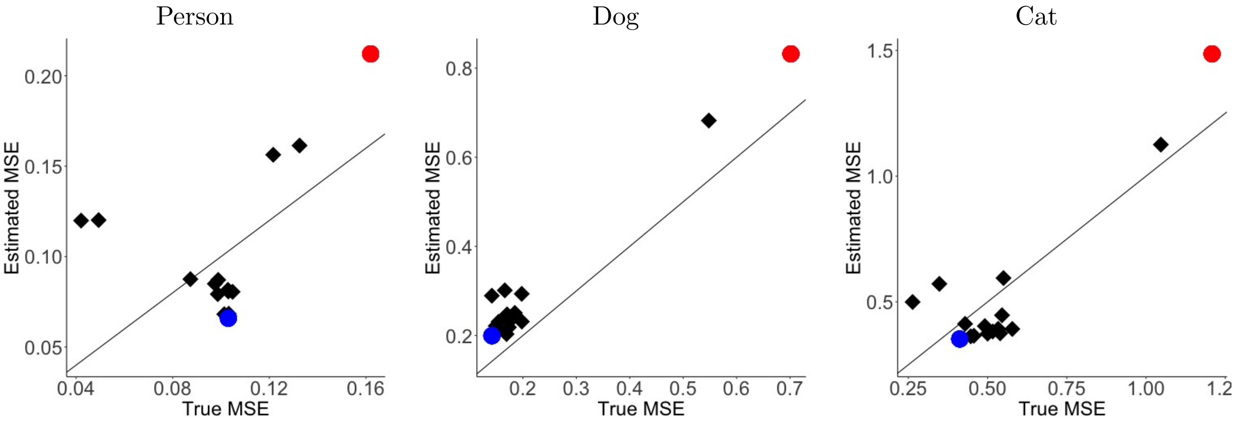

Figure 6: The bounding box experiment, true vs. estimated MSEs for the various class objects. The best and worst regressors according to SU-PCR algorithm are marked with green and red colors respectively.

Table 2: Average MSE of the different ensemble methods, MSE of the single best regressor, based on the estimated MSE and SU-PCR hit-rate, for the various prediction tasks. Averages and standard deviations (in parentheses) are computed over 20 train-test random splits. The Affairs dataset is not shown, as its estimated value $\hat{g}_2$ is very close to zero (see Fig. 2), namely accurate prediction by UPCR is not feasible for this dataset.

Dataset	Oracle	IU-PCR	SU-PCR	mean	median	Hit-Rate	$\arg\min_i \widehat{\text{MSE}_i}$
abalone	0.431 (0.006)	0.480 (0.009)	0.476 (0.011)	0.489 (0.007)	0.490 (0.010)	0.95	0.497 (0.013)
basketball	0.281 (0.005)	0.356 (0.006)	0.346 (0.005)	0.352 (0.004)	0.362 (0.004)	1.00	0.330 (0.018)
blog feedback	0.414 (0.025)	0.514 (0.013)	0.491 (0.016)	0.504 (0.015)	0.582 (0.017)	1	0.531 (0.111)
CCPP	0.059 (0.001)	0.066 (0.002)	0.065 (0.003)	0.068 (0.003)	0.065 (0.002)	0.75	0.068 (0.003)
CO2	0.445 (0.080)	0.752 (0.118)	0.737 (0.114)	0.763 (0.094)	0.812 (0.117)	0.75	0.812 (0.120)
crime-1	0.042 (0.017)	0.185 (0.108)	0.183 (0.102)	0.171 (0.101)	0.326 (0.185)	0.40	0.351 (0.203)
crime-2	0.084 (0.036)	0.185 (0.060)	0.199 (0.073)	0.177 (0.069)	0.301 (0.146)	0.35	0.378 (0.107)
flights AUS	0.328 (0.035)	0.613 (0.053)	0.611 (0.034)	0.582 (0.061)	0.664 (0.083)	0.55	0.598 (0.124)
flights BOS	0.469 (0.042)	0.677 (0.031)	0.633 (0.038)	0.659 (0.032)	0.689 (0.084)	0.85	0.620 (0.161)
flights BWI	0.439 (0.064)	0.742 (0.021)	0.673 (0.025)	0.707 (0.029)	0.822 (0.076)	0.95	0.623 (0.204)
flights HOU	0.396 (0.093)	0.712 (0.031)	0.630 (0.044)	0.688 (0.029)	0.753 (0.079)	1	0.569 (0.153)
flights JFK	0.495 (0.051)	0.777 (0.037)	0.714 (0.039)	0.744 (0.030)	0.898 (0.035)	1	0.819 (0.207)
flights LGA	0.471 (0.039)	0.731 (0.027)	0.663 (0.040)	0.704 (0.028)	0.778 (0.085)	0.95	0.653 (0.192)
flights longhaul	0.680 (0.047)	0.866 (0.031)	0.842 (0.074)	0.861 (0.059)	0.967 (0.013)	0.90	0.907 (0.138)
online videos	0.093 (0.006)	0.213 (0.015)	0.188 (0.012)	0.222 (0.014)	0.276 (0.020)	1	0.341 (0.062)
qsar	0.466 (0.028)	0.551 (0.031)	0.553 (0.033)	0.556 (0.029)	0.543 (0.040)	0.35	0.570 (0.048)
wine	0.595 (0.010)	0.658 (0.009)	0.642 (0.011)	0.660 (0.009)	0.687 (0.008)	1	0.624 (0.040)

Table 3: MSE of the different ensemble methods for the four HPN-DREAM prediction tasks.

Protein	Oracle	IU-PCR	SU-PCR	mean	median
UACC812	0.237	0.316	0.300	0.329	0.311
MCF7	0.300	0.424	0.396	0.508	0.472
BT549	0.193	0.264	0.253	0.283	0.302
BT20	0.141	0.201	0.215	0.212	0.208

Table 4: MSE of the different ensemble methods for the Person class of the bounding box task.

Coordinate	Oracle	IU-PCR	SU-PCR	mean	median
x_1	0.029	0.057	0.057	0.056	0.091
y_1	0.065	0.141	0.137	0.140	0.222
x_2	0.029	0.059	0.059	0.059	0.095
y_2	0.056	0.093	0.091	0.097	0.145

Table 5: MSE of the different ensemble methods for the Dog class of the bounding box task.

Coordinate	Oracle	IU-PCR	SU-PCR	mean	median
x_1	0.080	0.099	0.098	0.098	0.126
y_1	0.069	0.091	0.091	0.091	0.119
x_2	0.078	0.102	0.101	0.100	0.134
y_2	0.069	0.089	0.088	0.090	0.109

Table 6: MSE of the different ensemble methods for the Cat class of the bounding box task.

Coordinate	Oracle	IU-PCR	SU-PCR	mean	median
x_1	0.176	0.300	0.284	0.298	0.399
y_1	0.163	0.291	0.280	0.277	0.393
x_2	0.136	0.228	0.225	0.229	0.316
y_2	0.138	0.206	0.205	0.218	0.308

Table 7: Properties of interest of the datasets are shown in this table. n is the number of held-out samples. The input X is d dimensional, and the same n_{train} random samples were used to train the different algorithms in the ensemble.

Name	n	n_{TRAIN}	d
Abalone	3277	700	7
Affairs	5466	700	7
Basketball	48899	900	9
Blog Feedback	24197	28000	28055
CCPP	8968	400	4
CO2	84	40	5
Crime-1	325	150	127
Crime-2	325	150	127
Flights AUS	47595	1000	10
Flights BOS	112705	1000	10
Flights BWI	101665	1000	10
Flights HOU	53044	1000	10
Flights JFK	113960	1000	10
Flights LGA	111911	1000	10
Flights longhaul	9393	1000	10
Online Videos	66484	2100	21
qsar aquatic toxicity	547	150	8
Wine Quality White	3598	1100	11

Table 8: Parameters of regressors in manually crafted ensembles

Type	**Parameters**
Ridge	$\alpha = 0.5$
Kernel Regression	kernel chosen using cross validation between polynomial, RBF, sigmoid
Lasso	$\alpha = 0.1$
Orthogonal Matching Pursuit	-
Linear SVR	$C = 1$
SVR	RBF kernel with C chosen using cross validation out of 0.01, 0.1, 1, 10
Regression Tree	depth 4
Regression Tree	infinite depth
Random Forest	m=100 trees
Bagging Regressor	-