Ensembles of Regularized Linear Models

Anthony Christidis$^{(1)}$, Laks V.S. Lakshmanan$^{(2)}$, Ezequiel Smucler$^{(1)}$
Ruben Zamar$^{(1)}$

(1) Department of Statistics, University of British Columbia

(2) Department of Computer Science, University of British Columbia

Abstract

We propose an approach for building ensembles of regularized linear models by optimizing an objective function that encourages sparsity within each model and diversity among them. Our procedure works on top of a given penalized linear regression estimator (e.g., Lasso, Elastic Net, SCAD) by fitting it to possibly overlapping subsets of features, while at the same time encouraging diversity among the subsets, to reduce the correlation between the predictions from each fitted model. The predictions from the models are then aggregated. For the case of an Elastic Net penalty and orthogonal predictors, we give a closed form solution for the regression coefficients in each of the ensembled models. We prove the consistency of our method in possibly high-dimensional linear models, where the number of predictors can increase with the sample size. An extensive simulation study and real-data applications show that the proposed method systematically improves the prediction accuracy of the base linear estimators being ensembled. Possible extensions to GLMs and other models are discussed.

Keywords: high dimension small sample size; elastic net; lasso; linear regression

1 Introduction

Model ensembling is a powerful approach for prediction. Examples of ensemble methods for regression include Random Forests (Breiman, 2001) and Boosting (Schapire and Freund, 2012; Friedman, 2001). Both methods can adapt well to the presence of non-linearity, but the resulting prediction rules are generally difficult to interpret. If the relation between the response and the predictor variables is approximately linear, an ensemble of linear models will produce highly competitive predictions and yield more interpretable results.

We are interested in building ensembles of regularized linear models, with a special aim towards data sets in which the number of observations is relatively small, smaller or not much larger than the number of predictors. For motivation, consider the following toy example. We generate 500 replications of 10 independent observations from the model

$$y = 0x_1 + 1x_2 + 1x_3 + u,$$

where, x_i, $i = 1, \dots, 3$, and u are standard normal, u is independent of x_i, $i = 1, \dots, 3$, x_1 is independent of x_2 and x_3, and the correlation between x_2 and x_3 is 0.9. We fit three procedures to the data, the ordinary least squares estimator (LS), Elastic Net (EN) (Zou and Hastie, 2005) with penalty parameter chosen by leave-one-out cross-validation, and the following ensemble: apply least squares to the data using only predictors x_1 and x_2, then apply least squares to the data using only predictor x_3 and average the predictions from these two fits. We computed the prediction mean squared error (PMSE) of each procedure on an independent test set of size five thousand. The resulting PMSEs of LS and EN are 1.74 and 2.09, respectively, whereas the PMSE for the ensemble, 1.33, is much smaller.

The intuitive idea is that, for problems with a number of observations n that is relatively low when compared to the number of predictors p, the increase in bias due to leaving out variables from some of the models is compensated by a double reduction in variance: (i) the reduction in variance in each of the linear models due the lower dimensionality and possibly lower multicollinearity and (ii) the reduction in variance due to the averaging of the resulting predictors. Indeed, in the example above, the mean variances of LS and the ensemble are 0.74 and 0.32 respectively, whereas the mean squared biases are 2.6×10^{-3} and 9.4×10^{-3}. In our toy example, since predictors x_2 and x_3 are highly correlated, it seems sensible to place them in separate models. Note that to build the ensemble, in particular, to choose how to group the variables, we used our knowledge of the data generating process, which is not available in practice.

In general, one could exhaustively search over all possible groupings of the variables into different models and choose the one with the lowest estimated prediction error (e.g. using cross-validation), but this is computationally unfeasible. For example, the number of possible splits of p features into two groups of sizes p_1 and p_2 plus a third group of p_3 left-out features ($p_1 + p_2 + p_3 = p$) is 3^p. In general, the number of possible ensembles of G models plus a group of p_{G+1} left-out-features is $(G+1)^p$. This number becomes much larger if we allow the variables to be shared by the different models. At this point we notice that, in the simpler case of selecting a single subset of features (where $G = 1$), the combinatorial problem of evaluating 2^p possible subsets can be bypassed by using greedy algorithms such as forward stepwise regression or penalized estimators. An appropriately tuned penalized estimator, e.g. the lasso, is able to automatically and optimally determine which variables are left out and the required level of shrinkage applied to the active variables. We will see that a penalization approach can also be adopted to deal with the $G > 1$ case.

Suppose we have n samples of training data $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{X} \in \mathbb{R}^{n \times p}$, where $\mathbf{y}$ is the response variable and $\mathbf{X}$ is a matrix collecting all the available p features from each of the n samples, and we want to build $G > 1$ linear models using the data. We propose to minimize a penalized objective function of the form:

$$O(\mathbf{y},\mathbf{X},\boldsymbol{\beta}^1,\ldots,\boldsymbol{\beta}^G)=\sum_{g=1}^{G}\left(\frac{1}{2n}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^g\|^2+p_{\lambda_S}(\boldsymbol{\beta}^g)+q_{\lambda_D,g}(\boldsymbol{\beta}^1,\ldots,\boldsymbol{\beta}^G)\right), \qquad (1)$$

where $\boldsymbol{\beta}^g\in\mathbb{R}^p$ is the vector of coefficients for model g, p_{λ_S} is a penalty function, encouraging sparsity within the models and $q_{\lambda_D,g}$ is another penalty function, encouraging diversity among the models. In this paper, we take p_{λ_S} to be the Elastic Net penalty

$$p_{\lambda_S}(\boldsymbol{\beta}^g)=\lambda_S\left(\frac{(1-\alpha)}{2}\|\boldsymbol{\beta}^g\|_2^2+\alpha\|\boldsymbol{\beta}^g\|_1\right),$$

where $\alpha\in[0,1]$ and

$$q_{\lambda_D,g}(\boldsymbol{\beta}^1,\ldots,\boldsymbol{\beta}^G)=\frac{\lambda_D}{2}\sum_{h\neq g}\sum_{j=1}^{p}|\beta_j^h\beta_j^g|.$$

In general, by appropriately choosing the penalty function $p_{\lambda_S}(\boldsymbol{\beta}^g)$ it is seen that our method generalizes penalized regression estimators, such as the Lasso, the Elastic Net and the SCAD, (Tibshirani, 1996; Zou and Hastie, 2005; Fan and Li, 2001), allowing for the selection of possibly overlapping subsets of features in such a way that variables that work well together end up in the same model, while at the same time encouraging diversity between the models, to reduce the correlation between the predictions resulting from each of them. This implies that, by appropriately choosing the tuning parameters the proposed method automatically and optimally decides: which variables are left out, how many models are required, the distribution of the active variables among the different models (with possible overlap) and the shrinkage applied to the active variables in each of the models.

There has been a vast production of work, both theoretical and practical, dealing with regularization in linear models. The task of reviewing this mass of work is daunting and beyond the scope of this paper. The interested reader can find excellent reviews in Bühlmann and van de Geer (2011) and Hastie et al. (2015). The main difference between our approach and the existing methodology is that the final output of our approach consists of a collection of regression models, whose predictions can be averaged or otherwise combined to produce a final prediction. Moreover, our procedure allows for the optimal choice of the model sizes and overlap to yield better predicitions.

The rest of this article is organized as follows. In Section 2 we study the properties of the minimizers of (1) in some simple but illustrative cases. We prove a consistency result for the proposed method in Section 3. In Section 4 we propose an algorithm to compute the proposed estimators, to choose their tuning parameters and to aggregate the predictions from the constructed models. In the simulation study included in Section 5 we compare the performance with regards to prediction accuracy of the proposed method against that of several competitors. We apply the procedures considered in the simulation study to real data-sets in Section 6. Finally, some conclusions and possible extensions are discussed in Section 7. Technical proofs and additional simulations results are provided in the supplementaly material for this article.

2 Forming ensembles of regularized linear models

Assume we have training data $\mathbf{y}' = (y_1, \ldots, y_n)$, $\mathbf{x}_i' = (x_{i,1}, \ldots, x_{i,p})$, $i = 1, \ldots, n$ standardized so that

$$\frac{1}{n}\sum_{i=1}^{n} x_{i,j} = 0, \quad \frac{1}{n}\sum_{i=1}^{n} x_{i,j}^2 = 1, \quad 1 \leq j \leq p, \quad \frac{1}{n}\sum_{i=1}^{n} y_i = 0, \quad \frac{1}{n}\sum_{i=1}^{n} y_i^2 = 1$$

Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ be the matrix with $\mathbf{x}_i'$ as rows and let $\mathbf{x}^j$ be its columns.

We consider ensembles defined as minimizers of the objective function given in (1), that is

$$\hat{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta}} O(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}),$$

where $\boldsymbol{\beta} \in \mathbb{R}^{p\times G}$ is the matrix with $\boldsymbol{\beta}^g$ as columns and $O(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}) = O(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}^1, \dots, \boldsymbol{\beta}^G)$. Hence, as mentioned in the introduction, we use the quadratic loss to measure the goodness of fit of each model, the Elastic Net penalty to regularize each of the models, and the penalty $q_{\lambda_D,g}(\boldsymbol{\beta}^1, \dots, \boldsymbol{\beta}^G)$ to encourage diversity among the models.

The problem of minimizing (1) can be posed as an 'artificial' multivariate linear regression problem. Let $\mathbf{Y} \in \mathbb{R}^{n\times G}$ be the matrix with the vector $\mathbf{y}$ repeated G times as columns. Then

$$O(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}) = \frac{1}{2n}\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_F^2 + \lambda_S\left(\frac{(1-\alpha)}{2}\|\boldsymbol{\beta}\|_F^2 + \alpha\|\boldsymbol{\beta}\|_1\right) + \frac{\lambda_D}{2}\left(\||\boldsymbol{\beta}|'|\boldsymbol{\beta}|\|_1 - \|\boldsymbol{\beta}\|_F^2\right),$$

where $\|\cdot\|_F$ is the Frobenius norm, $|\boldsymbol{\beta}|$ stands for taking the absolute value coordinate-wise and $\|\cdot\|_1$ is the sum of the absolute values of the entries of the matrix. It is seen that the diversity penalty term in a sense penalizes correlations between the different models. Further insights can be gained by analyzing the term corresponding to each model in (1) separately. Fix any $1 \leq g \leq G$. Then

$$\begin{aligned}
&\frac{1}{2n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^g\|^2 + \lambda_S\left(\frac{(1-\alpha)}{2}\|\boldsymbol{\beta}^g\|_2^2 + \alpha\|\boldsymbol{\beta}^g\|_1\right) + \frac{\lambda_D}{2}\sum_{h\neq g}\sum_{j=1}^{p}|\beta_j^h\beta_j^g| \\
&= \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^g\|^2 + \lambda_S\frac{(1-\alpha)}{2}\|\boldsymbol{\beta}^g\|_2^2 + \sum_{j=1}^{p}|\beta_j^g|(\lambda_S\alpha + \frac{\lambda_D}{2}\sum_{h\neq g}|\beta_j^h|) \\
&= \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^g\|^2 + \lambda_S\frac{(1-\alpha)}{2}\|\boldsymbol{\beta}^g\|_2^2 + \sum_{j=1}^{p}|\beta_j^g|w_{j,g},
\end{aligned}$$

where $w_{j,g} = (\lambda_S\alpha + \lambda_D/2\sum_{h\neq g}|\beta_j^h|)$. Hence, when looking at each model separately, we are solving an Elastic Net type problem, where the Lasso penalty has weights which depend on the solution itself. In particular, the coordinates most penalized in model g will be those that have large coefficients in the other models. Some intuition on the impact of using our diversity penalty can be obtained by considering an extreme situation in which there is only one variable and three models. In Figure 1 we show level surfaces of the full penalty term for $p = 1$, $G = 3$, $\alpha = 1$, $\lambda_S = 1$ and different values of λ_D. Hence the surfaces plotted are the solutions of

$$|\beta_1^1| + |\beta_1^2| + |\beta_1^3| + \lambda_D \left(|\beta_1^1\beta_1^2| + |\beta_1^1\beta_1^3| + |\beta_1^3\beta_1^2|\right) = 1.$$

We see that when λ_D is small, the surface is similar to the three-dimensional ℓ_1 ball. For larger values of λ_D the surface becomes highly non-convex, with peaks aligned with the axes, where there is only one model that is non-null.

The following proposition follows easily from the previous discussion.

Proposition 1. *For $\lambda_D = 0$, the optimal $\hat{\boldsymbol{\beta}}$ has columns equal to the Elastic Net estimator.*

Since $\lambda_D = 0$ is always considered as a candidate penalty parameter, see Section 4, if the optimal model (in the sense of minimal cross-validated prediction mean squared error) is a single Elastic Net, this will be the final output of our method.

For any $\lambda_S > 0$, $O(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}) \to \infty$ as $\|\boldsymbol{\beta}\| \to \infty$ and hence a global minimum of O exists. The objective function O is not a convex function of $\boldsymbol{\beta}$ if $\lambda_D > 0$, due to the non-convexity of $q_{\lambda_D,g}$. Moreover, if $\hat{\boldsymbol{\beta}}$ is a global minimizer of O, any permutation of its columns is also a global minimizer. Importantly, the objective function is convex (strictly if $\alpha < 1$) in each coordinate and in each group of coordinates $\boldsymbol{\beta}^g$, since the corresponding optimization problems are actually penalized least squares problems with a weighted Elastic Net penalty.

2.1 The case of orthogonal predictors

We derive a closed form solution for the minimizers of the objective function in the special case in which the predictors are orthogonal. We find that the closed form solution for the orthogonal case provides some insights into how the procedure works.

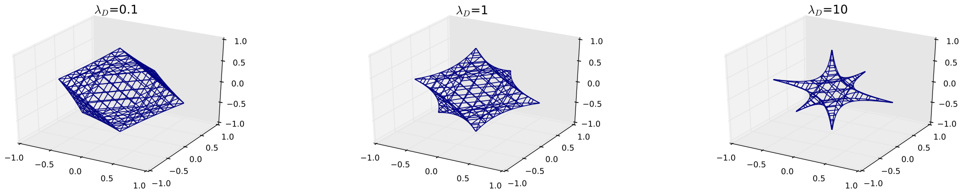

Figure 1: Plots of the full penalty term for $\alpha = 1$, $\lambda_S = 1$ and three different values of λ_D.

Proposition 2. *Assume $\mathbf{X}/\sqrt{n}$ is orthogonal and $G = 2$. Fix any $j = 1, \ldots, p$ and let $C_j = \mathbf{y}'\mathbf{x}^j/n$. Then*

1. *If $|C_j| \leq \alpha\lambda_S$ the j-th coefficients of all models in all solutions of the ensemble are zero.*

2. *If $|C_j| > \alpha\lambda_S$*

   (a) If $\lambda_D < 1 + (1-\alpha)\lambda_S$ all solutions of the ensemble satisfy

   $$\hat{\beta}_j^1 = \hat{\beta}_j^2 = \frac{\mathrm{soft}(C_j, \alpha\lambda_s)}{1 + (1-\alpha)\lambda_S + \lambda_D}.$$

   (b) If $\lambda_D = 1 + (1-\alpha)\lambda_S$ any pair (β_j^1, β_j^2) that satisfies $\beta_j^1\beta_j^2 \geq 0$ and

   $$\beta_j^1 + \beta_j^2 = \frac{\mathrm{soft}(C_j, \alpha\lambda_s)}{1 + (1-\alpha)\lambda_S}$$

   is a solution to the ensemble.

   (c) If $\lambda_D > 1 + (1-\alpha)\lambda_S$ all solutions of the ensemble satisfy that only one of $\hat{\beta}_j^1$ and $\hat{\beta}_j^2$ is zero, and the non-zero one is equal to

   $$\frac{\mathrm{soft}(C_j, \alpha\lambda_s)}{1 + (1-\alpha)\lambda_S}.$$

Some comments are in order. First, as happens with the classical Elastic Net, if the maximal correlation between the predictors and the response is smaller that $\alpha\lambda_S$, then all the coefficients in the ensemble are zero. Else, we have three distinct regimes. Fix some coordinate j. When $\lambda_D < 1 + (1-\alpha)\lambda_S$, the coefficients for predictor j in both models are equal, and are equal to the univariate Elastic Net estimator corresponding to penalty parameters α and λ_S but with an added ℓ_2 shrinkage: the factor dividing the soft thresholding operator has an added λ_D. If $\lambda_D = 1+(1-\alpha)\lambda_S$ the objective function depends only on $\beta_j^1 + \beta_j^2$ and hence more than one solution exists. Finally, if $\lambda_D > 1 + (1-\alpha)\lambda_S$, for all possible solutions of the ensemble, and for predictor j, only one of the models is non-null, and it is equal to the univariate Elastic Net.

2.2 The case of two correlated predictors

Further insights into how our procedure works can be gained by analyzing the simple case in which there are only two correlated predictors and two models.

Proposition 3. *Assume $\mathbf{X} \in \mathbb{R}^{n\times 2}$ is normalized so that its columns have squared norm equal to n and $G = 2$. Let $\hat{\boldsymbol{\beta}}$ be any solution of the ensemble, $\rho = (\mathbf{x}^2)'\mathbf{x}^1/n$ and $C_j = \mathbf{y}'\mathbf{x}^j/n$, $j = 1, 2$.*

1. *If the models are disjoint then the active variables in each model have coefficients*

$$T_j = \frac{\text{soft}(C_j, \alpha\lambda_s)}{1 + (1-\alpha)\lambda_S}, \quad j = 1, 2,$$

and

$$\lambda_D \geq \max\left\{\frac{|C_1 - \rho T_2| - \alpha\lambda_S}{T_1}, \frac{|C_2 - \rho T_1| - \alpha\lambda_S}{T_2}\right\}.$$

2. *If variable i is inactive in both models, variable j is active in both models and $\lambda_D \neq 1 + (1-\alpha)\lambda_S$ then the coefficients of variable j are equal to*
$$\frac{\mathrm{soft}(C_j, \alpha\lambda_s)}{1 + (1-\alpha)\lambda_S + \lambda_D}.$$

3. *Assume $\lambda_S = 0$ and that both variables are active in both models. If* $\mathrm{sign}(\hat{\beta}_1^1) = \mathrm{sign}(\hat{\beta}_1^2)$ *and* $\mathrm{sign}(\hat{\beta}_2^1) = \mathrm{sign}(\hat{\beta}_2^2)$ *then all solutions of the ensemble satisfy*
$$\begin{pmatrix} 1 & \lambda_D & 0 & \rho \\ \lambda_D & 1 & \rho & 0 \\ 0 & \rho & 1 & \lambda_D \\ \rho & 0 & \lambda_D & 1 \end{pmatrix} \begin{pmatrix} \hat{\beta}_1^1 \\ \hat{\beta}_1^2 \\ \hat{\beta}_2^2 \\ \hat{\beta}_2^1 \end{pmatrix} = \begin{pmatrix} C_1 \\ C_1 \\ C_2 \\ C_2 \end{pmatrix}.$$
If $\lambda_D < 1 - \rho$, the solution is unique.

The case in which $\lambda_S = 0$ is easier to analyze. In this case, the proposition above implies that if the fitted models are disjoint then $\lambda_D \geq \{|1 - \rho(C_1/C_2)|, |1 - \rho(C_2/C_1)|\}$, and the non-null coefficients in the ensemble are equal to the marginal Elastic Net regressions. Note that in the case in which $C_1 = C_2$, the size of the diversity penalty required to separate the models decreases as the correlation between the variables increases.

3 A consistency result

Assume the data follows a standard linear regression model
$$y_i = \mathbf{x}_i'\boldsymbol{\beta}_0 + \varepsilon_i, \quad 1 \leq i \leq n, \tag{2}$$
where the vector of predictors $\mathbf{x}_i$ is fixed and the errors ε_i are i.i.d. normal random variables with variance σ^2. The number of predictors p may depend on the sample size and be greater than n. As before, we assume that $(1/n)\sum\limits_{i=1}^{n} x_{i,j}^2 = 1$ for $j = 1 \ldots p$.

Theorem 1. *Assume that* $\lambda_S \geq \sigma\sqrt{(t^2 + 2\log(p))/n}$ *for some* $t > 0$*. Let* $\hat{\boldsymbol{\beta}}$ *be any solution of the ensemble, with* $\alpha = 1$*. Then with probability at least* $1 - 2\exp(-t^2/2)$ *we have*

$$\frac{1}{2n}\left\| \left(\frac{1}{G}\sum_{g=1}^{G} \mathbf{X}\hat{\boldsymbol{\beta}}^g\right) - \mathbf{X}\boldsymbol{\beta}_0 \right\|_2^2 \leq 2\lambda_S \|\boldsymbol{\beta}_0\|_1 + \frac{\lambda_D(G-1)}{2}\|\boldsymbol{\beta}_0\|_2^2.$$

It follows that if we take λ_S to be of order $\sqrt{\log(p)/n}$ and λ_D to be of order $\log(p)/n$ then if we assume $\|\boldsymbol{\beta}_0\|_1$ is of order smaller than $\sqrt{n/\log(p)}$ and $\log(p)/n \to 0$, the average prediction of the ensemble, $(1/G)\sum_{g=1}^{G} \mathbf{X}\hat{\boldsymbol{\beta}}^g$, is consistent. A similar result can be obtained if one assumes only that the errors have a sub-gaussian distribution. Sharper bounds may be obtained if one assumes more restrictive conditions on the set of predictor variables, for example the so-called *compatibility condition*; see Section 6.2.2 of Bühlmann and van de Geer (2011) for details. An overview of consistency results for regularized estimators is available in, for example, Bühlmann and van de Geer (2011). In the more classical case in which the smallest eigenvalue of $\mathbf{X}'\mathbf{X}/n$ is bounded below by a fixed constant, for example when p is taken to be fixed and $\mathbf{X}'\mathbf{X}/n$ converges to a positive definite matrix, we may deduce that $\|(1/G)\sum_{g=1}^{G} \hat{\boldsymbol{\beta}}^g - \boldsymbol{\beta}_0\|_2^2 \to 0$ in probability as n goes to infinity.

4 Algorithm

4.1 Computing solutions for fixed penalty parameters

To obtain approximate solutions of the minimizer of (1), we propose an algorithm based on coordinate descent: we cycle through the coordinates of $\boldsymbol{\beta}$, optimizing with respect to each coordinate while keeping the others fixed. Coordinate descent has proven to be very efficient in solving regularized least squares problems, see Friedman et al. (2010) for example.

Proposition 4. *The coordinate descent update for β_j^g is*

$$\beta_j^{n,g} = \frac{\text{soft}\left(\frac{1}{n}\sum\limits_{i=1}^{n} x_{i,j}(y_i - y_i^{(-j),g}), \alpha\lambda_S + \lambda_D \sum\limits_{h\neq g} |\beta_j^{o,h}|\right)}{1 + (1-\alpha)\lambda_S},$$

where $y_i^{(-j),g}$ is the in-sample prediction of y_i using model g and leaving out variable j, soft is the soft-thresholding operator, defined by soft(z,γ) = sign$(z)\max(0, |z| - \gamma)$, the superscript n stands for the new solution and the superscript o stands for the old solution.

The proof of Proposition 4 is straightforward and for this reason it is ommited. Note that the ℓ_1 shrinkage being applied to variable j in model g, $\alpha\lambda_S + \lambda_D \sum_{h\neq g} |\beta_j^{o,h}|$, increases with the sum of the absolute values of the coefficients of variable j in all other models. This shows more clearly that the penalty $(\lambda_D/2)\sum_{h\neq g}\sum_{j=1}^{p} |\beta_j^h \beta_j^g|$ encourages diversity among the models.

We cycle through the coordinates of $\boldsymbol{\beta}^1$, then through those of $\boldsymbol{\beta}^2$ and so on until we reach $\boldsymbol{\beta}^G$, where we check for convergence. Convergence is declared when

$$\max_j \left(\frac{1}{G}\sum_{g=1}^{G} \beta_j^{n,g} - \frac{1}{G}\sum_{g=1}^{G} \beta_j^{o,g}\right)^2 < \delta,$$

for some small positive δ. Since the data is standardized, the convergence criterion in the original units is:

$$\max_j \frac{1}{n}\sum_{i=1}^{n} \left(x_{i,j}\frac{1}{G}\sum_{g=1}^{G} \beta_j^{n,g} - x_{i,j}\frac{1}{G}\sum_{g=1}^{G} \beta_j^{o,g}\right)^2 < \delta\frac{1}{n}\sum_{i=1}^{n} (y_i - \bar{y})^2.$$

Hence, the algorithm converges when the in-sample average predictions no longer change significantly. If the algorithm did not converge, we start over.

Remark 1. *It follows easily from Theorem 4.1 of Tseng (2001) that the proposed algorithm converges to a coordinate-wise minimum of* (1).

4.2 Aggregating the predictions

Once we have computed the G models $\hat{\boldsymbol{\beta}}^1, \ldots, \hat{\boldsymbol{\beta}}^G$, we aggregate them to form a predictor by averaging the models: if $\mathbf{x}$ is a new observation the prediction of the response is

$$\widehat{y}(\mathbf{x}) = \frac{1}{G} \sum_{g=1}^{G} \mathbf{x}' \hat{\boldsymbol{\beta}}^g = \mathbf{x}' \left(\frac{1}{G} \sum_{g=1}^{G} \hat{\boldsymbol{\beta}}^g \right) = \mathbf{x}' \hat{\boldsymbol{\beta}}_*, \tag{3}$$

where $\hat{\boldsymbol{\beta}}_* = (1/G) \sum_{g=1}^{G} \hat{\boldsymbol{\beta}}^g$, which is an estimate of the regression coefficients. Theorem 1 is a consistency result for this way of averaging the models.

Breiman (1996) proposes to aggregate predictors by averaging them according to weights determined by solving a constrained least squares problem and calls the method *stacking*. In detail, given predictors $v_k(\mathbf{x}), k = 1, \ldots, K$, define their leave-one-out versions as $v_k^{-i}(\mathbf{x}), i = 1, \ldots, n$. Let $z_{k,i} = v_k^{-i}(\mathbf{x}_i)$. Then the weights used to form the final predictor are defined as the non-negative constants $\alpha_1, \ldots, \alpha_K$ that minimize $\sum_{i=1}^{n} \left(y_i - \sum_{k=1}^{K} \alpha_k z_{k,i} \right)^2$. Breiman also provides empirical evidence to show that using 10-fold cross-validation instead of leave-out-out to generate the data can be more effective, as well as less computationally demanding.

The theoretical properties of combining prediction procedures are discussed in Yang (2004) and references therein.

4.3 Choosing the penalty parameters

We choose λ_S and λ_D over grids of candidates, looking to minimize the cross-validated (CV) mean squared error (MSE). The grids of candidates are built as follows. It is easy to show that, for $\lambda_D = 0$ and $\alpha > 0$, the smallest λ_S that makes all the models null is given by $\lambda_S^{max} = 1/(n\alpha) \max_{j \leq p} |\sum_{i=1}^{n} x_{i,j} y_i|$. λ_S^{max} is the maximum sparsity penalty that will be considered. The smallest λ_D that maximises diversity among the models (makes them disjoint) for a given λ_S, say λ_D^{max}, is estimated using a grid search. Proposition 3 hints that in general λ_D^{max} will depend in a complicated way on the correlations between the predictors. To build a grid to search for the optimal λ_S we take 100 log-equispaced points between $\varepsilon\lambda_S^{max}$ and λ_S^{max}, where ε is 10^{-4} if $p < n$ and 10^{-2} otherwise. The grid used for λ_D is built analogously, but including zero as a candidate.

Even though we could also cross-validate over a grid of possible values of α, we find that taking a large value of α, say $\alpha = 3/4$ or $\alpha = 1$, generally works well and hence in what follows we assume that α is fixed.

Fix one of λ_S, λ_D. We then minimize the objective function O over the grid of candidates corresponding to the other penalty term, going from the largest to the smallest values in the grid; for each element of the grid of candidates, the solution to the problem using the previous element is used as a warm start. Even though the optimal $\hat{\boldsymbol{\beta}}$ is not in general a continuous function of λ_D and λ_S, see Proposition 2, we find that using warm starts as described above works well in practice.

The main loop of the algorithm works as follows, starting with $\lambda_D^{opt} = 0$, and until the CV MSE no longer decreases:

- Find the λ_S in the grid giving minimal CV MSE, λ_S^{opt}.
- Take the optimal λ_S^{opt} from the previous step. Recompute λ_D^{max} and the corresponding grid. Find the λ_D in the grid giving minimal cross-validated MSE, λ_D^{opt}. Go to the previous step.

As we mentioned earlier, since we start with $\lambda_D^{opt} = 0$, the solution with all columns equal to the Elastic Net estimator is always a candidate to be chosen.

4.4 Choosing the number of models

We conduct a small simulation study to illustrate the effect increasing the number of models has on the computation time and the performance of an ensemble of Lassos. We generate 100 replications of a linear model with $p = 1000$ predictors and $n = 100$ observations, corresponding to the second covariance structure described in Section 5. For each replication, two data-sets are generated, one on which the ensemble is trained and one used for computing the prediction mean squared error (PMSE). The computation is repeated for various values of the proportion of active variables, called ζ. The signal to noise ratio is 10. We show the PMSEs for different values of the number of models used (rows) and the proportion of active variables in the data generating process (columns). We also computed a measure of the overlap between the models in the ensemble. Let $\hat{\boldsymbol{\beta}} \in \mathbb{R}^{p\times G}$ be the matrix with columns equal to the computed models, where G is the number of models and p the number of features. Let $o_j = (1/G)\sum_{g=1}^{G} I\{\hat{\beta}_j^g \neq 0\}$, then we define the overlap as

$$\text{OVP} = \frac{\sum\limits_{j=1}^{p} o_j I\{o_j \neq 0\}}{\sum\limits_{j=1}^{p} I\{o_j \neq 0\}}$$

if $\sum_{j=1}^{p} I\{o_j \neq 0\} \neq 0$, and as 0 otherwise. Note that $0 \leq \text{OVP} \leq 1$. If $\text{OVP} = 0$ then all models are empty, whereas if $\text{OVP} > 0$, then at least one model is non-empty and actually $\text{OVP} \geq 1/G$. If $\text{OVP} = 1/G$ then each variable that is active can only appear in one model, and hence the overlap between the models is minimal, since they are disjoint. Finally, if $\text{OVP} = 1$ then all the variables that are active in at least one model, actually appear in all the models, and hence we have maximum overlap.

Table 1 shows the results. The last column shows the average computation time in seconds. The computation time doesn't vary much between different sparsity levels, and hence we report the average over them. In this case, as the number of models used increases, both the overlap and the PMSE decrease, but the gain in prediction accuracy due to using more models also decreases. There seems to be a 'diminishing returns' type phenomenon. Of course, this pattern may not persist in other settings. An objective way to determine the number of models to be used, is to cross-validate over a coarse grid, say, taking $2, 5, 7$ or 10 models; this is the approach we take in Section 6, where we apply the proposed methodology to a real data-set. In all the settings studied in this paper, the increase in computational time due to using more models, appears to be approximately linear in the number of models, as evidenced by Table 1. In our simulations we always use ten models, a possibly sub-optimal choice, but still good enough to give a excellent performance.

	$\zeta = 0.1$		$\zeta = 0.2$		$\zeta = 0.3$		
	PMSE	OVP	PMSE	OVP	PMSE	OVP	Time
2	1.21	0.62	1.18	0.61	1.17	0.60	17.91
5	1.16	0.37	1.13	0.36	1.12	0.35	39.37
7	1.15	0.33	1.12	0.29	1.12	0.32	52.00
10	1.15	0.32	1.11	0.26	1.10	0.27	70.30

Table 1: PMSEs, overlap and average computation time in seconds for different values of the number of models (rows) and proportion of active variables ζ (columns) for SNR=10.

5 Simulations

5.1 Methods

We ran a simulation study, comparing the prediction accuracy of the following nine competitors. All computations were carried out in `R`.

1. The **Lasso**, computed using the `glmnet` package.

2. The **Elastic Net** with $\alpha = 3/4$, computed using the `glmnet` package.

3. An ensemble of Lassos, using $G = 10$ models, called **Ens-Lasso**.

4. An ensemble of Elastic Nets, with $\alpha = 3/4$, using $G = 10$ models, called **Ens-EN**.

5. The sure independence screening (SIS) procedure, Fan and Lv (2008), followed by fitting a SCAD penalized least squares estimator, computed using the `SIS` package, called **SIS-SCAD**.

6. The MC+ penalized least squares estimator, Zhang (2010), computed using the `sparsenet` package, called **SparseNet**.

7. The Relaxed Lasso, Meinshausen (2007), computed using the `relaxnet` package, called **Relaxed**.

8. The forward stepwise algorithm, computed using the `lars` package, called **Stepwise**.

9. The Cluster Representative Lasso, proposed in Bühlmann et al. (2013), computed using code kindly provided by the authors, called **CRL**.

10. The Random Forest of Breiman (2001), computed using the `randomForest` package, called **RF**.

11. The Random GLM method of Song et al. (2013), computed using the `randomGLM` package, called **RGLM**.

All tuning parameters were chosen via cross-validation; the RF and RGLM methods were ran using their default settings. The CRL of Bühlmann et al. (2013) was not included in scenarios with $p = 1000$ due to its long computation time when compared with the rest of the methods. For the same reason, in the scenarios with $p = 150$, we only did 100 replications for CRL, instead of the 500 done for all the other procedures.

The popular Group Lasso (Yuan and Lin, 2006; Simon et al., 2013) is not included in the simulation, because we don't assume that there is a priori knowledge of the existence of pre-defined groups among the features. The interesting recent proposals of Bühlmann et al. (2013), Sharma et al. (2013) and Witten et al. (2014), assume that there exist unknown clusters of correlated variables, and shrink the coefficients of variables in the same cluster towards each other. Because CRL has a relatively more efficient numerical implementation (compared with the other two) we included it in our simulation to represent the cluster-based approaches. Finally, note that all the competitors above, except perhaps for the forward stepwise algorithm, the RF and the RGLM, could in principle be used as building blocks in our procedure for forming ensembles. Hence, our main objective in this simulation study is to show that the proposed method for building ensembles improves upon the prediction accuracy of the base estimators being ensembled, in this case, the Lasso and the Elastic Net.

5.2 Models

For each Monte Carlo replication, we generate data from a linear model:

$$y_i = \mathbf{x}_i' \boldsymbol{\beta}_0 + \sigma \epsilon_i, \quad 1 \leq i \leq n,$$

where the $\mathbf{x}_i \in \mathbb{R}^p$ are multivariate normal with zero mean and correlation matrix $\mathbf{\Sigma}$ and the ϵ_i are standard normal. We consider different combinations of p and n, see below. For each p we take the number of active variables to be $p_0 = [p\zeta]$ for $\zeta = 0.05, 0.1, 0.2, 0.3$ and 0.4. Given p, n and a sparsity level $1-\zeta$, the following scenarios for $\boldsymbol{\beta}_0$ and $\mathbf{\Sigma}$ are considered

Scenario 1 $\Sigma_{i,j} = \rho$ for all $i \neq j$, the first $[p\zeta]$ coordinates of $\boldsymbol{\beta}_0$ are equal to 2 and the rest are 0.

Scenario 2

$$\Sigma_{i,j} = \begin{cases} 1 & \text{if } i = j \\ \rho & \text{if } 1 \leq i,j \leq \lfloor p_0/2 \rfloor + \lceil (p-p_0)/2 \rceil, i \neq j \\ \rho & \text{if } \lfloor p_0/2 \rfloor + \lceil (p-p_0)/2 \rceil + 1 \leq i,j \leq p, i \neq j \\ 0 & \text{otherwise} \end{cases}$$

$\beta_j = 1$ for $j \leq \lfloor p_0/2 \rfloor$, $\beta_j = -1$ for $\lfloor p_0/2 \rfloor + \lceil (p-p_0)/2 \rceil + 1 \leq j \leq \lfloor p_0/2 \rfloor + \lceil (p - p_0)/2 \rceil + p_0 - \lfloor p_0/2 \rfloor$ and the rest of the coordinates equal to zero.

Scenario 3

$$\Sigma_{i,j} = \begin{cases} 1 & \text{if } i = j \\ \rho & \text{if } 1 \leq i,j \leq p_0, i \neq j \\ 0 & \text{otherwise,} \end{cases}$$

the first $[p\zeta]$ coordinates of $\boldsymbol{\beta}_0$ are equal to 2 and the rest are 0.

We consider different values of ρ: 0.2, 0.5, 0.8. Then σ is chosen to give a desired signal to noise ratio (SNR), defined as

$$\text{SNR} = \frac{\boldsymbol{\beta}_0' \mathbf{\Sigma} \boldsymbol{\beta}_0}{\sigma^2}.$$

We consider SNRs of 3, 5, 10 and 50. In the first scenario all the predictors are correlated among each other. In scenario two we have two groups of active variables. This is similar to the simulation scenario considered in Witten et al. (2014). Variables within each group are correlated with each other, but the groups are independent. In the third scenario the active variables are only correlated with each other. We report results for Scenario 1 with $(p, n) = (1000, 100)$ and $\rho = 0.2$, Scenario 2 with $(p, n) = (150, 75)$ and $\rho = 0.8$ and with $(p, n) = (1000, 100)$ and $\rho = 0.5$, and Scenario 3 with $(p, n) = (1000, 50)$ and $\rho = 0.5$.

5.3 Performance measures

For each replication, two independent copies of the data are generated, one to fit the procedures, the other one to compute the prediction mean squared error (PMSE), divided by the variance of the noise, σ^2. Hence, the best possible result is 1. In each table reporting the PMSEs we also compute the standard error for each of the methods, and report the maximum among them in the caption.

We also compute the precision (PR) and recall (RC) of each method, defined as

$$\mathrm{PR} = \frac{\#\{j : \beta_{0,j} \neq 0 \wedge \beta_j \neq 0\}}{\#\{j : \beta_j \neq 0\}}, \quad \mathrm{RC} = \frac{\#\{j : \beta_{0,j} \neq 0 \wedge \beta_j \neq 0\}}{\#\{j : \beta_{0,j} \neq 0\}}.$$

For the ensembles, the vector of coefficients used to compute the precision and recall is the average of the models, see (3). For the SIS-SCAD method, the precision and recall are computed using the variables selected by the SIS step. For the RGLM method, the precision and recall are computed using the union of the variables selected in each of the bags. Since RF does not fit a linear model, we do not compute its precision and recall.

5.4 Results

In the scenarios we consider the Elastic Net and the Lasso have very similar behaviours, as do the Ensemble of Elastic Nets and the Ensemble of Lassos. For this reason, and due to the fact that we are comparing eleven different procedures, we will only report the results for: the Lasso, the ensemble of Lassos, and the best performing among the remaining methods, excluding the Lasso, the Elastic Net and the two ensembles. The full results of the simulation can be found in the Supplement for this article.

Tables 2 to 5 show the results, which can be summarized as follows. The ensemble does as well or better than the base Lasso. In cases with SNR $= 3$, the improvements generally range from around 5 to 15%, whereas in cases with SNR $= 5, 10$ improvements range from around 10 to 30%. In cases with an admittedly high SNR of 50, the ensemble can have a PMSE that is half or less than half that of the base estimator. In general, as expected, the improvements tend to increase with the SNR and with the proportion of active variables. Moreover, in the majority of the cases considered here, the ensemble has the lowest PMSE of all the competitors considered. In Scenarios 1 and 3, the strongest competitor is RGLM, with PMSEs similar to that of the ensemble for SNRs of 3 and 5. However, for higher SNRs the ensemble tends to have a better performance. In Scenario 2 the strongest competitors are CRL and SparseNet, with performances similar to that of the ensemble for $\zeta = 0.05, 0.1$. For $\zeta = 0.2, 0.3, 0.4$, the ensemble tends to have a lower PMSE. It is important to note that in Scenario 2, for $p = 150, n = 75$, the RGLM has a rather poor performance, with PMSEs that can be double those of the ensemble; see Table 19 in the Supplement for example.

We also note that in general the recall of the ensemble is higher than that of the base estimator and those of the other competitors, except the RGLM. The price to pay for this improvement is a decrease in precision, generally minor, but in some cases important (for example in Table 5).

		$\zeta = 0.05$			$\zeta = 0.1$			$\zeta = 0.2$			$\zeta = 0.3$			$\zeta = 0.4$		
SNR		PMSE	RC	PR	PMSE	RC	PR	PMSE	RC	PR	PMSE	RC	PR	PMSE	RC	PR
3	Lasso	1.55	0.14	0.12	1.45	0.1	0.16	1.41	0.08	0.24	1.38	0.07	0.33	1.37	0.07	0.43
	Ens-Lasso	1.35	0.47	0.09	1.25	0.38	0.13	1.2	0.36	0.23	1.18	0.34	0.32	1.17	0.34	0.42
	RGLM	1.29	0.79	0.06	1.2	0.75	0.11	1.15	0.72	0.21	1.14	0.71	0.31	1.13	0.7	0.41
5	Lasso	1.78	0.19	0.13	1.63	0.13	0.17	1.55	0.1	0.25	1.52	0.09	0.34	1.5	0.09	0.44
	Ens-Lasso	1.49	0.57	0.09	1.33	0.49	0.13	1.25	0.46	0.23	1.21	0.45	0.32	1.19	0.44	0.42
	RGLM	1.49	0.81	0.06	1.35	0.76	0.11	1.28	0.72	0.21	1.25	0.72	0.31	1.24	0.71	0.41
10	Lasso	2.29	0.26	0.15	2.03	0.17	0.18	1.89	0.13	0.27	1.84	0.11	0.35	1.81	0.1	0.44
	Ens-Lasso	1.83	0.69	0.09	1.53	0.63	0.13	1.34	0.61	0.23	1.27	0.61	0.32	1.24	0.6	0.42
	RGLM	2.02	0.82	0.06	1.78	0.77	0.11	1.64	0.74	0.21	1.6	0.72	0.31	1.58	0.72	0.41
50	Lasso	6.76	0.36	0.19	6.54	0.2	0.21	6.36	0.14	0.29	6.26	0.12	0.37	6.24	0.11	0.45
	Ens-Lasso	4.46	0.86	0.09	3.09	0.86	0.13	2.44	0.83	0.23	2.23	0.81	0.33	2.13	0.8	0.42
	RGLM	6.54	0.83	0.06	5.4	0.78	0.11	4.81	0.74	0.21	4.6	0.73	0.31	4.49	0.73	0.41

Table 2: Mean PMSEs, recalls and precisions for Scenario 1 with ρ =0.2, n =100, p =1000. Maximum standard error is 0.05.

		$\zeta = 0.05$			$\zeta = 0.1$			$\zeta = 0.2$			$\zeta = 0.3$			$\zeta = 0.4$		
SNR		PMSE	RC	PR	PMSE	RC	PR	PMSE	RC	PR	PMSE	RC	PR	PMSE	RC	PR
3	Lasso	1.37	0.1	0.1	1.3	0.07	0.14	1.27	0.06	0.24	1.27	0.06	0.33	1.26	0.06	0.42
	Ens-Lasso	1.24	0.35	0.08	1.18	0.29	0.12	1.15	0.28	0.22	1.14	0.26	0.32	1.14	0.26	0.42
	RGLM	1.26	0.75	0.05	1.2	0.72	0.11	1.18	0.7	0.2	1.17	0.7	0.3	1.16	0.69	0.4
5	Lasso	1.52	0.13	0.11	1.41	0.09	0.15	1.36	0.08	0.24	1.35	0.07	0.34	1.35	0.07	0.43
	Ens-Lasso	1.33	0.44	0.08	1.23	0.38	0.13	1.18	0.35	0.22	1.17	0.34	0.32	1.15	0.34	0.42
	RGLM	1.41	0.77	0.06	1.32	0.73	0.11	1.28	0.71	0.2	1.27	0.7	0.3	1.26	0.7	0.4
10	Lasso	1.84	0.19	0.13	1.65	0.13	0.16	1.55	0.1	0.25	1.52	0.1	0.35	1.5	0.09	0.44
	Ens-Lasso	1.53	0.58	0.09	1.34	0.52	0.13	1.24	0.48	0.23	1.21	0.47	0.32	1.19	0.47	0.42
	SparseNet	1.83	0.24	0.14	1.63	0.17	0.18	1.54	0.14	0.29	1.51	0.13	0.39	1.5	0.12	0.49
50	Lasso	4.43	0.32	0.17	4.19	0.18	0.19	4.02	0.13	0.27	3.98	0.12	0.36	3.95	0.11	0.45
	Ens-Lasso	2.95	0.83	0.09	2.13	0.8	0.13	1.74	0.76	0.23	1.63	0.74	0.33	1.56	0.73	0.43
	SparseNet	3.94	0.45	0.2	3.65	0.29	0.24	3.52	0.21	0.34	3.49	0.18	0.44	3.45	0.16	0.52

Table 3: Mean PMSEs, recalls and precisions for Scenario 2 with ρ =0.5, n =100, p =1000. Maximum standard error is 0.03.

		$\zeta = 0.05$			$\zeta = 0.1$			$\zeta = 0.2$			$\zeta = 0.3$			$\zeta = 0.4$		
SNR		PMSE	RC	PR	PMSE	RC	PR	PMSE	RC	PR	PMSE	RC	PR	PMSE	RC	PR
3	Lasso	1.3	0.5	0.18	1.26	0.28	0.2	1.23	0.2	0.28	1.21	0.18	0.37	1.2	0.17	0.45
	Ens-Lasso	1.27	0.74	0.12	1.2	0.61	0.16	1.16	0.54	0.25	1.14	0.53	0.34	1.13	0.51	0.43
	CRL	1.27	0.51	0.18	1.26	0.27	0.19	1.23	0.2	0.27	1.21	0.18	0.37	1.2	0.17	0.45
5	Lasso	1.38	0.66	0.21	1.34	0.36	0.22	1.29	0.25	0.3	1.27	0.22	0.38	1.25	0.2	0.46
	Ens-Lasso	1.35	0.82	0.15	1.25	0.72	0.16	1.19	0.64	0.25	1.16	0.62	0.34	1.15	0.6	0.44
	CRL	1.34	0.66	0.21	1.34	0.35	0.22	1.28	0.25	0.29	1.27	0.22	0.38	1.26	0.2	0.46
10	Lasso	1.49	0.85	0.24	1.48	0.5	0.26	1.42	0.33	0.32	1.37	0.28	0.4	1.35	0.26	0.49
	Ens-Lasso	1.5	0.91	0.19	1.37	0.81	0.17	1.25	0.77	0.25	1.2	0.75	0.34	1.17	0.74	0.44
	CRL	1.44	0.85	0.24	1.48	0.49	0.25	1.4	0.33	0.33	1.37	0.28	0.41	1.35	0.26	0.48
50	Lasso	1.55	1	0.28	2.01	0.86	0.34	2.02	0.59	0.4	1.92	0.49	0.47	1.86	0.43	0.54
	Ens-Lasso	1.56	1	0.28	2	0.92	0.26	1.7	0.91	0.26	1.49	0.93	0.34	1.37	0.94	0.42
	SparseNet	1.34	0.99	0.62	1.95	0.89	0.34	1.97	0.65	0.42	1.89	0.55	0.5	1.83	0.49	0.58

Table 4: Mean PMSEs, recalls and precisions for Scenario 2 with ρ =0.8, n =75, p =150. Maximum standard error is 0.02.

		$\zeta = 0.05$			$\zeta = 0.1$			$\zeta = 0.2$			$\zeta = 0.3$			$\zeta = 0.4$		
SNR		PMSE	RC	PR	PMSE	RC	PR	PMSE	RC	PR	PMSE	RC	PR	PMSE	RC	PR
3	Lasso	1.5	0.27	0.65	1.43	0.16	0.72	1.38	0.1	0.75	1.35	0.07	0.79	1.33	0.06	0.81
	Ens-Lasso	1.44	0.5	0.59	1.35	0.41	0.64	1.27	0.31	0.68	1.24	0.24	0.73	1.21	0.21	0.74
	RGLM	1.5	1	0.06	1.36	1	0.14	1.25	1	0.37	1.17	1	0.82	1.18	0.96	0.95
5	Lasso	1.61	0.33	0.69	1.53	0.2	0.76	1.45	0.12	0.8	1.42	0.09	0.82	1.4	0.07	0.85
	Ens-Lasso	1.5	0.59	0.61	1.4	0.5	0.65	1.29	0.39	0.69	1.26	0.31	0.73	1.23	0.27	0.76
	RGLM	1.63	1	0.06	1.42	1	0.14	1.27	1	0.38	1.18	1	0.87	1.17	0.96	0.97
10	Lasso	1.8	0.42	0.73	1.72	0.26	0.8	1.61	0.16	0.87	1.57	0.12	0.9	1.54	0.09	0.92
	Ens-Lasso	1.6	0.7	0.63	1.46	0.63	0.65	1.33	0.5	0.7	1.29	0.42	0.74	1.26	0.36	0.78
	RGLM	1.91	1	0.06	1.54	1	0.14	1.31	1	0.38	1.22	1	0.92	1.2	0.97	0.99
50	Lasso	2.75	0.63	0.83	2.81	0.4	0.95	3.08	0.23	1	3.34	0.16	1	3.49	0.12	1
	Ens-Lasso	1.89	0.91	0.69	1.69	0.83	0.74	1.5	0.78	0.77	1.47	0.74	0.81	1.48	0.7	0.85
	RGLM	3.88	1	0.06	2.35	1	0.14	1.67	1	0.39	1.53	1	0.95	1.7	0.97	0.99

Table 5: Mean PMSEs, recalls and precisions for Scenario 3 with ρ =0.5, n =50, p =1000. Maximum standard error is 0.06.

6 Glass data-sets

We analyze the performance of the competitors considered in the previous section when predicting on real data-sets from a chemometric problem. To evaluate the prediction accuracy of the competitors we randomly split the data into a training set that has 50% of the observations and a testing set that has the remaining 50%. This is repeated 100 times and the resulting prediction MSEs are averaged. The results are reported relative to the best average performance among all estimators. Hence, the estimator with the best average performance will have a score of 1. We also report the average rank among the data-sets for each method.

The glass data-sets (Lemberge et al., 2000) were obtained from an electron probe X-ray microanalysis (EPXMA) of archaeological glass samples. A spectrum on 1920 frequencies was measured on a total of 180 glass samples. The goal is to predict the concentrations of the following chemical compounds using the spectrum: Na2O, MgO, Al2O3, SiO2, P2O5, SO3, Cl, K2O, CaO, MnO, Fe2O3, BaO and PbO. After removing predictors with little variation, we are left with $p = 486$ frequencies and $n = 180$ observations. The CRL estimator was not included in the comparison, due to its long computation time. The number of models used to form the ensembles is chosen by cross-validation among the values 2, 5, 7, 10. The Elastic Net and the Ensemble of Elastic Nets were computed with $\alpha = 0.1$, closer to a Ridge than a Lasso estimator, since we a priori expected a relatively low level of sparsity.

Table 6 shows the results. Highlighted in black is the best performing method for each compound. It is seen that the Ensemble of Lassos has the best overall behavior, having the highest average rank (1.92) over the thirteen compounds. Excluding the Ensemble of Elastic Nets which performs similarly to the Ensemble of Lassos, the RGLM is the strongest competitor, with a rank of 3.54.

	Na2O	MgO	Al2O3	SiO2	P2O5	SO3	Cl	K2O	CaO	MnO	Fe2O3	BaO	PbO	Rank
Lasso	1.70	1.16	1.38	2.78	1.19	1.14	1.39	1.41	1.48	1.09	1.38	1.04	1.01	4.23
Elastic Net	1.62	1.18	1.74	2.73	1.31	1.17	1.13	1.44	1.47	1.05	1.23	1.01	1.08	4.62
Ens-Lasso	1.31	**1.00**	**1.00**	1.67	**1.00**	**1.00**	**1.00**	1.36	1.40	1.05	1.26	1.03	**1.00**	**1.92**
Ens-EN	1.58	1.18	1.66	2.56	1.25	1.12	1.10	1.43	1.47	1.04	1.20	**1.00**	1.08	3.38
SparseNet	1.63	1.31	1.76	2.84	1.28	1.18	1.54	1.53	1.57	1.15	1.42	1.06	1.06	6.31
Relaxed	1.80	1.23	1.40	2.81	1.16	1.19	1.38	1.40	1.49	1.13	1.46	1.10	1.28	5.62
Stepwise	2.40	2.13	3.72	4.10	2.65	1.52	4.75	1.89	1.78	1.18	1.51	2.32	1.73	9.00
RF	**1.00**	1.95	7.04	3.02	16.48	1.16	12.06	6.93	9.18	1.16	**1.00**	2.69	6.34	7.77
RGLM	1.80	1.02	1.20	**1.00**	**1.00**	1.26	1.62	**1.00**	**1.00**	**1.00**	1.19	1.72	1.04	3.54
SIS-SCAD	2.01	4.19	1.70	2.88	1.85	1.29	1.87	2.08	2.08	1.19	1.67	1.68	1.95	8.62

Table 6: Average PMSEs for each compound over 100 random splits into training and testing sets. Last column shows the average rank over all compounds.

7 Discussion

We have proposed a novel method for forming ensembles of linear regression models. Examples using real and synthetic data-sets show that the approach systematically improves the prediction accuracy of the base estimators being ensembled. In the synthetic data-sets, the improvements tend to increase with the signal to noise ratio and the number of active variables. We believe that the results reported in this paper show that the proposed method is a valuable addition to the practitioners toolbox.

The approach taken in this paper can be extended in several ways. Other sparsity penalties such as the SCAD can be handled similarly. In fact, the algorithm proposed here will work with any regularized model approach provided the coordinate descent updates can be expressed in closed form. Our method can be extended to GLMs by ensembling regularized GLM estimators instead of linear regression estimators. For example, ensembles of logistic regression models can be formed by replacing the quadratic loss in (1) with the deviance. The method can be robustified to deal with outliers by using, for example, a bounded loss function to measure the goodness of fit of each model in (1), instead of the classical least squares loss; in this case regularized robust regression estimators (see Smucler and Yohai (2017) for example) would be ensembled. Lower computational times may be achieved by using early stopping strategies when computing solution paths over one of the penalties and also by using an active set strategy when cycling over the groups, see Friedman et al. (2010).

SUPPLEMENTARY MATERIAL

The supplemental material available online contains the proofs of the theoretical results stated in the paper and the full results of our simulation study. An `R` package that implements the procedures proposed in this paper, called `ensembleEN` is available from `CRAN`.

References

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

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

Bühlmann, P., Rütimann, P., van de Geer, S., and Zhang, C.-H. (2013). Correlated variables in regression: Clustering and sparse estimation. *J. Statist. Planng Inf.*, 143(11):1835 – 1858.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Series in Statistics. Springer Berlin Heidelberg.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Ass.*, 96(456):1348–1360.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Statist. Soc. B*, 70(5):849–911.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Ann. Statist.*, 29(5):1189–1232.

Friedman, J. H., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman & Hall/CRC.

Lemberge, P., De Raedt, I., Janssens, K. H., Wei, F., and Van Espen, P. J. (2000). Quantitative analysis of 16-17th century archaeological glass vessels using pls regression of epxma and μ-xrf data. *Journal of Chemometrics*, 14(5-6):751–763.

Meinshausen, N. (2007). Relaxed lasso. *Computational Statistics & Data Analysis*, 52(1):374–393.

Schapire, R. E. and Freund, Y. (2012). *Boosting: Foundations and Algorithms*. The MIT Press.

Sharma, D. B., Bondell, H. D., and Zhang, H. H. (2013). Consistent group identification and variable selection in regression with correlated predictors. *Journal of Computational and Graphical Statistics*, 22(2):319–340.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245.

Smucler, E. and Yohai, V. J. (2017). Robust and sparse estimators for linear regression models. *Computational Statistics & Data Analysis*, 111(C):116–130.

Song, L., Langfelder, P., and Horvath, S. (2013). Random generalized linear model: a highly accurate and interpretable ensemble predictor. *BMC bioinformatics*, 14(1):5.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B*, 58(1):267–288.

Tseng, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *J. Optim. Theory Appl.*, 109(3):475–494.

Witten, D., Shojaie, A., and Zhang, F. (2014). The cluster elastic net for high-dimensional regression with unknown variable grouping. *Technometrics*, 56(1):112–122.

Yang, Y. (2004). Aggregating regression procedures to improve performance. *Bernoulli*, 10(1):25–47.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc. B*, 68(1):49–67.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, 38(2):894–942.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Statist. Soc. B*, 67(2):301–320.