Technical University of Denmark

Selecting groups of covariates in the elastic net

Clemmensen, Line Katrine Harder

Publication date:
2014

Document Version
Publisher's PDF, also known as Version of record

Citation (APA):
Clemmensen, L. K. H. (2014). Selecting groups of covariates in the elastic net. Kgs. Lyngby: Technical University of Denmark (DTU). (DTU Compute-Technical Report-2014; No. 15).

Selecting groups of covariates in the elastic net

Line H. Clemmensen
Department of Applied Mathematics and Computer Science
Technical University of Denmark
Lyngby, Denmark

DTU Compute Technical Report-2014-15

August 22, 2014

Abstract

This paper introduces a novel method to select groups of variables in sparse regression and classification settings. The groups are formed based on the correlations between covariates and ensure that for example spatial or spectral relations are preserved without explicitly coding for these. The preservation of relations gives increased interpretability. The method is based on the elastic net and adaptively selects highly correlated groups of variables and does therefore not waste time in grouping irrelevant variables for the problem at hand. The method is illustrated on a simulated data set and on regression of moisture content in multispectral images of sand. In both cases, the predictions were better or similar to existing regression and classification algorithms and the interpretation was enhanced using the grouping method. On top of that, the grouping method more consistently selects the important variables.

I. Introduction

Highly correlated covariates are common in for example image analysis where correlations exist both spatially and spectrally. The correlations arise naturally since the measurements are taken from an underlying continuous process. The two facts: that data come from an underlying continuous process and that there are many similar (highly correlated) variables in the high-dimensional data, were described by Donoho (2000) as blessings of dimensionality. This implies two things. One, that the underlying structure of the data often is simple, i.e. that data lie on a low-dimensional manifold in the high-dimensional sampling space. Two, that we can average over the highly correlated variables and thereby obtain more robust estimates by averaging out noise in data.

This paper exploits these blessings of dimensionality to obtain better predictions and better interpretation when we have highly-correlated variables which are to be used in either regression or classification. We take a starting point in the sparse regression approach called the elastic net (Zou and T.Hastie, 2005).

The elastic net regularizes the ordinary least squares solution with both an L_1- and an L_2-norm constraint on the parameter estimates. The elastic net combines the good convergence rate of the L_1-norm and the small error estimates from the L_2-norm (Jin et al., 2009). Firstly, the sparseness induced by the L_1-norm helps finding a low-dimensional manifold on which data can be represented. Secondly, the L_2-norm shrinks the parameter estimates and thereby averages over highly correlated variables to obtain more generalizable results. However, the elastic net does not provide a framework to include the highly correlated variables into the model. This paper proposes a way to include highly correlated groups of variables into the elastic net model and thereby obtain better interpretation and prediction.

Previously, most work on grouping variables together aims at making the groups of covariates in one step and then making a prediction model in a second step using for example varimax rotated principal components or in other ways predefining the groups before applying classification or regression techniques (Yuan and Lin, 2006; Wei and Huang, 2010). The proposed method adaptively makes a group in each iteration of the least angle regression selection (LARS, Efron et al. (2004)), which the elastic net algorithm builds on. This way, we don't waste time in grouping variables that are irrelevant to the task at hand (the relevant predictions), but concentrate on grouping variables that are relevant for the low-dimensional manifold we are seeking. Reccently, similar approaches have been taken in Bondell and Reich (2008), Shen and Huang (2010), and Shara et al. (2013).

The Grouping pursuit by Shen and Huang (2010) involves a penalty involving pairwise comparisons of all parameter estimates $\beta_j - \beta_{j'}$. The Grouping pursuit does not include variable selection into the model and thus differs from the proposed method in sparsity and in a grouping based on the parameter estimates rather than the correlations between the variables. Looking only at a subset of selected variables makes the grouping more time-efficient in particular for problems with a high number of variables where we expect only a subset of these to be of importance for the problem at hand.

The OSCAR (Octogonal Shrinkage and Clustering Algorithm for Regression) method by Bondell and Reich (2008) adds an L_1- and a pairwise L_∞-norm penalty to the parameters. The pairwise infinity norm encourages every pair of parameters to be of equal size. The OSCAR method encourages sparseness like the proposed method, but differs in that the grouping based on the parameter estimates rather than the correlations between the variables. We here note that the grouping property of giving parameter estimates equal size when the variables are correlated arise from the L_2-norm and thus we would not expect the approaches to be equivalent as no L_2-norm is added to the OSCAR model. This algorithms work well for low-dimensional problems ($p \simeq 20$), but are not applicable for high-dimensional problems. The more recent PACS (Pairwise Absolute Clustering and Sparisty) method by Bondell and Reich (2008) groups variables by penalizing the L_1-norm of both the absolute differences and sums of the parameter estimates. Additionally, the L_1-norm of the paraeter estimates is added to obtain varaible selection. PACS includes a discussion of only including pairs of variables with correaltions above some threshold. They also comment that "We notice that the PACS approaches do not perform as well in prediction and selection as the existing selection approaches and that the elastic-net approaches perform the best in terms of prediction and selection". For this reason we will limit our comparison to the elastic net.

Other methods dealing with the problem of uncertainty in the variable selection of high dimensional problems have been proposed, such as tilted correlation screenign (TCS) and tilted correaltion screenign learning (TCSL) (Cho and Fryzlewicz, 2012; Lin and Pang, 2013). These methods deal with the uncertainty in the correlation estimates when $p \gg n$, and focuses on a correct estimation of the correlation between the predictors and the response by taking into accoutn the correlations amongst the covariates. There is no grouping performed in these studies, but for further researhc it would be of interest to consider tilted correlations in the setting proposed here.

The rest of the paper is organized as follows. Section two reviews the elastic net regression model and describes the shortcomings regarding selection of highly correlated variables. Subsequently, we propose the group elastic net algorithm, which rather than selecting single variables, selects groups of highly correlated variables. Section three contains experimental results and a comparison of the group elastic net and the elastic net on a synthetic data set, and one example with image data. The discussion is in section three, and we conclude in section 4.

II. Methodology

This section briefly explains the elastic net (Zou and T.Hastie, 2005), the least angle regression selection (Efron et al., 2004) and the group lasso (Yuan and Lin, 2006), and then uncovers the problem of selecting highly correlated variables and grouping these. Finally, an algorithm for adaptively grouping highly correlated covariates is proposed.

I. Elastic net

Zou and Hastie proposed the elastic net in 2005 (Zou and T.Hastie, 2005). The elastic net minimizes the sum of squared errors while penalizing the size of the L_1- and L_2-norm of the parameter estimates. The model parameters are obtained as

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} \{ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \|\boldsymbol{\beta}\|_2^2 \} \quad , \tag{1}$$

where $\mathbf{X}$ is an $n \times p$ matrix of n observations with p variables, $\mathbf{y}$ is an $n \times 1$ vector of the measured output, and $\boldsymbol{\beta}$ is a $p \times 1$ vector of the parameter estimates. Here, $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|$, $|\cdot|$ denoting the absolute value, and $\|\boldsymbol{\beta}\|_2^2 = \sum_{j=1}^{p} \beta_j^2$. Choosing $\lambda_1 = 0$ yields ridge solutions (Hoerl and Kennard, 1970), and likewise choosing $\lambda_2 = 0$ yields lasso solutions (Tibshirani, 1996).

The elastic net algorithm augments the data such that the L_2 constraint is fulfilled, and then using an algorithm called the least angle regression and selection (LARS; Efron et al. (2004)) solves the L_1-penalized problem. For further details see Zou and T.Hastie (2005). The LARS algorithm works on centered and standardized data, i.e.

$$\sum_{i=1}^{n} y_i = 0, \quad \sum_{i=1}^{n} x_{ij} = 0, \quad \sum_{i=1}^{n} x_{ij}^2 = 1, \; j = 1, ..., p. \tag{2}$$

This means that the inner product of the covariates, $r_{lj} = \mathbf{x}_l^T \mathbf{x}_j$ is the correlation between the l^{th} and the j^{th} variable. In each step LARS takes a step along the variable with maximum absolute correlation to the current residual. The correlation of variable j and the residual is given by

$$c_j = \mathbf{x}_j^T (\mathbf{y} - \boldsymbol{\mu}_k) \ , \tag{3}$$

where $\boldsymbol{\mu}_k$ is the estimated prediction at step k. Early stopping in form of selecting $k < p$ is used as equivalent to λ_1 for obtaining sparseness. In the first step LARS takes a step in the direction of the variable with the maximum correlation $C = \max |c_j|$. It proceeds until another variable has an equal correlation to the residual (or an equiangular angle) as those variables already included in the model (also called the active set). LARS includes the variable with an equal correlation and proceeds in the direction of the equiangular vector $\mathbf{u}_{\mathcal{A}}$ of the active set $\mathcal{A}$, given by:

$$\mathbf{u}_{\mathcal{A}} = \mathbf{X}_{\mathcal{A}} \mathbf{w}_{\mathcal{A}}, \tag{4}$$

where the covariates of the active set of variables is denoted as $\mathbf{X}_{\mathcal{A}}$, and $\mathbf{w}_{\mathcal{A}} = A_{\mathcal{A}} \mathbf{X}_{\mathcal{A}}^T \mathbf{X}_{\mathcal{A}} \mathbf{1}_{\mathcal{A}}$, with $A_{\mathcal{A}} = (\mathbf{1}_{\mathcal{A}}^T \mathbf{X}_{\mathcal{A}}^T \mathbf{X}_{\mathcal{A}} \mathbf{1}_{\mathcal{A}})^{-1/2}$, and $\mathbf{1}_{\mathcal{A}}$ is a vector of ones with the size of the active set $\mathcal{A}$. The length of each step is given by:

$$\gamma = \min_{j \in \mathcal{I}} \left\{ \frac{C - c_j}{A_{\mathcal{A}} - (\mathbf{X}_{\mathcal{I}}^T \mathbf{u}_{\mathcal{A}})_j}, \frac{C + c_j}{A_{\mathcal{A}} + (\mathbf{X}_{\mathcal{I}}^T \mathbf{u}_{\mathcal{A}})_j} \right\}, \tag{5}$$

where $\mathcal{I} = \mathcal{A}^c$ is the set of inactive variables and complementary to the set of active variables, and thus $\mathbf{X}_{\mathcal{I}}$ denotes the covariates for the inactive set of variables. The update of the estimate is then

$$\boldsymbol{\mu}_{k+1} = \boldsymbol{\mu}_k + \gamma \mathbf{u}_{\mathcal{A}}, \tag{6}$$

and with the regression coefficients given as

$$\boldsymbol{\beta}_{k+1} = \gamma(\boldsymbol{\beta}_{OLS}^{\mathcal{A}} - \boldsymbol{\beta}_k) + \boldsymbol{\beta}_k, \tag{7}$$

where $\boldsymbol{\beta}_{OLS}^{\mathcal{A}} = (\mathbf{X}_{\mathcal{A}}^T \mathbf{X}_{\mathcal{A}})^{-1} \mathbf{X}_{\mathcal{A}}^T \mathbf{y}$ is the OLS solution based on the variables in $\mathcal{A}$.

In the original LARS algorithm one variable is added to $\mathcal{A}$ at each iteration, and the number of iterations or non-zero elements can be used to control the L_1-norm penalty. The higher λ_1 in (1), the lower the number of iterations, and likewise non-zero elements in the model. This is also referred to as *early stopping* as there is in general no need to let the LARS algorithm include all variables and converge to the ordinary least squares solution.

The elastic net can be used for classification by regressing on indicator dummy variables of 0s and 1s for each of the classes. For a more thorough way of obtaining a sparse classification based on the elastic net, the reader is referred to sparse discriminant analysis (SDA) (Clemmensen et al., 2011). SDA provides a low-dimensional subspace smaller than the number of classes. In these settings (regression on indicator variables or SDA), the grouping can also be implemented. The sparse discriminant analysis additionally expands to nonlinear separation by use of Gaussian mixtures.

II. Predefined groups in lasso

A group lasso algorithm was developed in Yuan and Lin (2006) where predefined groups were selected into the least angle regression selection (LARS) model. This group lasso uses the mean value of the squared correlations between the variables in each group and the current residual as an estimate of the most correlated set. The current most correlated set is then given by

$$\mathcal{G} = \mathrm{argmax}_g \|\mathbf{X}_g^T(\boldsymbol{y} - \boldsymbol{\mu}_{\mathcal{A}})\|_2^2 / p_g, \tag{8}$$

where p_g is the number of variables in group g, and $\mathbf{X}_g$ is the set of covariates in group g. In Winham et al. (2011) it is shown that selecting such predefined groups of covariates rather than individual variables gives an increased generalization power. However, in general we do not have predefined groups available.

III. Adaptive grouping of variables

This section outlines the ideas behind using an adaptive grouping of variables. The adaptive grouping algorithm based on the elastic net is given in the following section.

III.1 Covariates with equivalent correlations with the current residual

First note that the correlation between the dependent variable $\mathbf{y}$ and an independent variable $\mathbf{x}_j$ is $c_j = \mathbf{x}_j^T\mathbf{y}$ since $\mathbf{y}$ and $\mathbf{x}_j$ are both centered and normalized to unit length.

Zou and Hastie mention in their paper on the elastic net, that: *The elastic net has the ability to do grouped selection*; (Zou and T.Hastie, 2005, p. 315). They refer to the method assigning almost identical coefficients to strongly correlated variables. However, this assumes that the strongly correlated variables are all selected into the model. Variables which are strongly correlated do not get selected in one step in their algorithm. First one of the variables $\mathbf{x}_j$ gets selected. In the following step, it is not likely that the strongly correlated variable, $\mathbf{x}_l$ will get selected. This results from the partial correlation between $\mathbf{y}$ and $\mathbf{x}_l$ conditioned on $\mathbf{x}_j$ becoming smaller than the unconditioned correlation when $\mathbf{x}_j$ and $\mathbf{x}_l$ are correlated

$$\mathrm{Corr}[\mathbf{y}, \mathbf{x}_l | \mathbf{x}_j] \propto \mathrm{Corr}[\mathbf{y}, \mathbf{x}_l] - \mathrm{Corr}[\mathbf{y}, \mathbf{x}_j]\mathrm{Corr}[\mathbf{x}_l, \mathbf{x}_j] \quad . \tag{9}$$

This can likewise be seen by returning to the LARS algorithm and rewriting (3) with only one variable in the active set to

$$c_j = \mathbf{x}_j^T \mathbf{y} - \mathbf{x}_j^T(\gamma \mathbf{X}_{\mathcal{A}}) \ . \tag{10}$$

We note that correlation c_j increases when variable j has a large correlation with the output y, but decreases when variable j has a large correlation with the active variable $\mathbf{X}_{\mathcal{A}}$, again illustrating that highly correlated variables are not favored in the least angle regression algorithm, nor the elastic net algorithm.

Therefore, we need to modify the algorithm to select groups of variables rather than individual variables. Note, that once the highly correlated variables have entered into the model, the grouping is performed through the ℓ_2-penalization assigning almost identical coefficients to the highly correlated variables. The fact that two or more variables are entered as a group in one iteration is not of concern. The partial correlation between $\mathbf{y}$ and $\mathbf{x}_l$ conditioned on $\mathbf{x}_j$ becomes larger than the unconditioned correlation when $\mathbf{x}_j$ and $\mathbf{x}_l$ are uncorrelated, cf. (9). Hence, as in the original algorithm, if $\mathbf{x}_l$ is uncorrelated with the variables already included in the model, but correlated with the output, then $\mathbf{x}_l$ will enter in the following iteration.

We consider variables with roughly equal sized correlations with the dependent variable to enter in one iteration. That is, for an observed (fixed) maximal correlation size $|c_i|$, we consider for the correlation c_j and some small constant δ that

$$if \quad c_j - \delta \leq c_i \leq c_j + \delta \quad \Rightarrow \quad c_j \equiv c_i \quad , \tag{11}$$

i.e. we accept that c_j equals c_i.

III.2 Correlations among covariates

We have now established how we can consider two covariates to have equivalently sized correlations with our current residual. Then we turn to defining concise groups of covariates within such covariates. We compute the correlations between the covariates of consideration, usually a subset which is much smaller than p. Only covariates with correlations higher than some threshold r_t are grouped together. In order to minimize the number of parameters in the algorithm, and as δ and r_t naturally are related, we choose to only have one parameter and set $\delta = 1 - r_t$. The threshold depends on the data problem at hand, but we have found values around $r_t \in [0.7, 0.95]$ to be suitable for the types of problems we have considered where high correlations exist.

IV. Proposed grouping algorithm

The algorithm proposed here is used within the elastic net framework, but could easily be adapted to other sequential variable selection algorithms (sometimes also called path algorithms). The adaptive grouping algorithm is

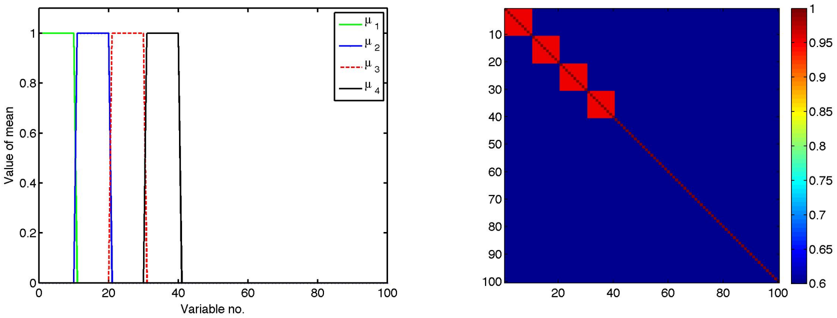
Algorithm 1 Group elastic net algorithm

```
1. Require $\mathbf{X}$, $\mathbf{y}$, and $r_t$.
2. Ensure $\sum_i y_i = 0$, $\sum_i x_{ij} = 0$, $\sum_i x_{ij}^2 = 1$, $\forall j$.
3. Initialize $\mathcal{A} = \emptyset$, $\mathcal{I} = \{1, ..., p\}$, $\delta = 1 - r_t$, $k = 0$ and $\boldsymbol{\mu}_k = \mathbf{0}$.
4. While early stopping criterion not met
   (a) Compute the current correlations $c_j = \mathbf{x}_j^T(\mathbf{y} - \boldsymbol{\mu}_k)$, $j \in \mathcal{I}$, with
       maximum corerlation $C = \max(\text{abs}(c_j))$ and corresponding index
       $\mathcal{M} = \arg\max_j(\text{abs}(c_j))$.
   (b) Identify the set of variables with correlations of equivalent size to $C$,
       as $\mathcal{P} = \text{find}(\text{abs}(c_j - C) \leq \delta)$ , $j \in \mathcal{I} \setminus \mathcal{M}$.
   (c) Compute the correlations between the variable with maximum current
       correaltion and all variables in the identified set $\mathcal{P}$, $\mathbf{r} = \mathbf{X}_{\mathcal{M}}^T \mathbf{X}_{\mathcal{P}}$.
       Find the the set of grouped covariates as the variables which have
       suitable correlation sizes $\mathcal{J} = find(\mathbf{r} > r_t)$.
   (d) Compute the step length $\gamma$ and the equiangular direction $\mathbf{u}_{\mathcal{A}}$ using
       (4) and (5), and update the current prediction $\boldsymbol{\mu}_{k+1} = \boldsymbol{\mu}_k + \gamma \mathbf{u}_{\mathcal{A}}$,
       and the set of active variables $\mathcal{A}_{k+1} = \mathcal{A}_k \cup \mathcal{M} \cup \mathcal{J}$.
   (e) End of iteration, $k = k + 1$.
5. Return the active set of variables and the predictions.
```

As early stopping we use the number of non-zero elements, i.e. we stop when we pass a number of non-zero elements n_z. n_z substitutes the parameter λ_1 in (1) and can be chosen using cross-validation on the training data; similarly the other parameters λ_2, and r_t can be chosen using cross-validation. For the purpose of examining how the group elastic net performs we will in the following illustrate the results for various values of r_t.

III. Results

The proposed group elastic net is illustrated on two different data sets. The first example consists of a simulated data set of four groups with 1000 variables and a correlation structure with strong correlations between the covariates. Here, elastic net regression on dummy variables was used and a linear discriminant analysis (LDA) was applied to the fitted values. The second example considers



(a) (b)

Figure 1: (a): The first 100 entries in the means of the four classes in the simulation study. The remaining entries in the means are all zero. (b): Correlation matrix of the first 100 variables in the simulation study. The remaining entries in the correlation matrix are all 0.6 in the off-diagonal and 1 in the diagonal.

prediction of moisture content in sand used to make concrete. The 2016 variables used for prediction were extracted from multispectral images of 59 sand samples (Clemmensen et al., 2010). Here we use regression of the continuous output, which is the measured reference moisture content.

I. Classification of simulated data

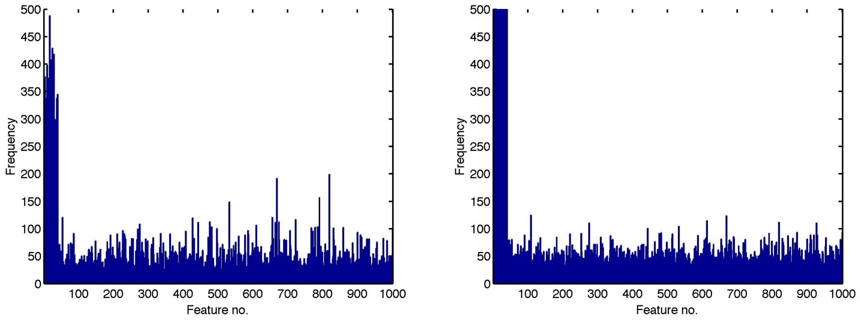
Consider a model with four Gaussian distributed classes $\sim N(\mu_j, \Sigma)$ of 1000 variables. The first 40 variables differ in mean for the four classes within groups of 10 variables. Additionally, the four groups of ten variables are highly correlated with a pair-wise correlation of 0.95. The remaining pair-wise correlations between the 1000 variables are 0.6. The means of the four classes are illustrated in Fig. 1(a) and the correlation matrix is illustrated in Fig. 1(b). Training data was simulated with 100 samples in each class (a total of 400 samples), and another 400 samples were simulated and used as test data. The simulations were performed using `mvnrnd` in Matlab and were repeated 50 times.

Five-fold cross-validation on the training data set was used to set parameters, and the separate test data set was used to estimate the classification rates. The results are summarized in Table 1. The group elastic net performed better with respect to classification rates and seems to overfit less than the elastic net classification. For each discriminative direction in GEN, four groups of variables were selected. The four groups selected by the group elastic net matched the true groups of variables. For the elastic net, approximately half of the selected variables were not the ones with true differences between the groups. In comparison, the group elastic net generally only selected the variables with true differences. For the optimal $r_t = 0.7$ the average number of non-zero variables was 37 which is close to the true number (40).

Table 1: Summary of average error rates using the elastic net (EN) and group elastic net (GEN) on the simulated data with a test set of 400 samples. The values are mean and (standard deviation) over 50 simulations.

Method	Training [%]	Test [%]	n_z	λ_2	r_t
EN	18.2 (3.7)	27.8 (4.3)	25.3 (14.4)	7800 (4000)	-
GEN	22.5 (2.6)	26.1 (2.4)	27.8 (14.4)	3900 (4600)	0.9
GEN	23 (3)	26.2 (2.7)	30.7 (14.9)	2500 (4000)	0.8
GEN	23.3 (3.1)	25.7 (2.1)	37.3 (12.5)	1500 (3200)	0.7
GEN	23.5 (2.6)	25.8 (2.3)	30.4 (15.3)	1700 (3400)	0.6

Fig. 2 shows the frequency of selection for each of the 1000 variables in 500 bootstrap samples of data. The group elastic net more consistently selected the true important variables (the first 40) than the elastic net.



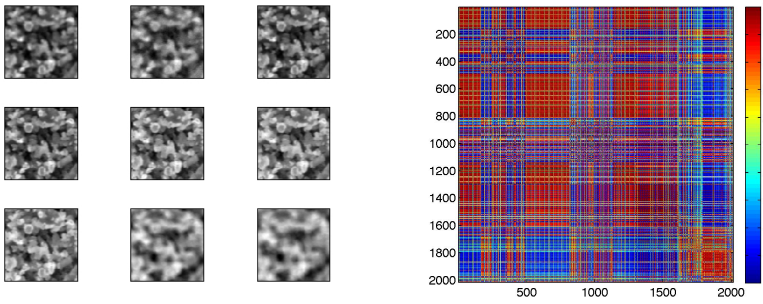
(a)

(b)

Figure 2: Selected variables for the elastic net with and without grouping over 500 bootstrappings of 100 samples each. Here, the stopping criterion used was $n_z = 25$ and $n_z = 37$ non-zero elements, respectively, which is the average of the optimal for (a): the elastic net, and (b): the group elastic net.

II. Predictions of moisture content based on multispectral images of sand

This study consists of 59 multispectral images of sand samples with varying moisture content (1-9% moisture). The multispectral images consist of nine spectral bands ranging from the visual to the near infrared area (428-940nm). 2016 summary statistics, such as the mean, standard deviation or percentiles of the intensities in a single spectral band or in pair-wise differences between spectral bands, were extracted from each multispectral image. The variables represent both the spatial and the spectral information in each sample, for more information see Clemmensen et al. (2010). The correlation matrix of the covariates is illustrated in Fig. 3. High correlations exist among the covariates.



(a) (b)

Figure 3: (a): Example of the 9 spectral bands of one of the sand images. (b): Correlation matrix of the variables extracted from the multispectral sand images for the training set.

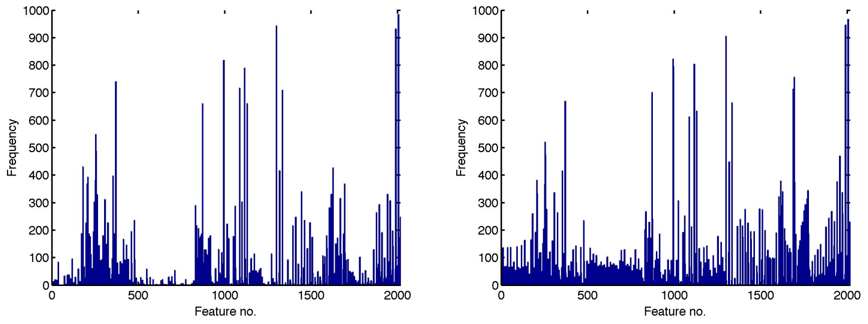
The samples were randomly split into a training set of 40 observations, and a test set of 19 observations. Leave-one-out cross validation was used on the training data set to select suitable parameters for the model, and then prediction errors were estimated using these parameters on the test set. This procedure was repeated 50 times. Table 2 summarizes the optimal parameter choices and the average estimated prediction errors for the elastic net (EN) and the group elastic net (GEN). Note, that it was optimal to include more variables in GEN than it was in EN.

Examining one of the selected models, the group elastic net had four groups of variables. In all four groups, the variables were close in index number. Close index numbers are in general related to summary statistics like the 90th and the 95th percentile of one of the spectral bands. We thus also saw in the correlation matrix (Fig. 3) blocks of variables next to each other with high correlations.

For the sand data we see that there is no statistical significant difference between the group elastic net and the elastic net in terms of test error, but the test error and its standard deviation were lower for the best group elastic net. Fig. 4 illustrates the frequency of selection for each of the 2016 variables for 1000 bootstrap samples of the sand data. There are differences in the selected variables, but it is hard to make any conclusions based on these, most likely due to the generally high correlations between most covariates.

Table 2: Summary of average prediction errors using elastic net (EN) and group elastic net (GEN) on the sand data. Resampling was performed 50 times; using 40 observations for training which included a leave-one-out cross validation, and 19 observations for testing.

Method	MSE Train	MSE Test	n_z	λ_2	r_t
EN	0.14 (0.15)	1.82 (2.1)	39.7 (27.1)	0.027 (0.035)	-
GEN	0.16 (0.14)	1.82 (1.57)	42.8 (27.8)	0.028 (0.033)	0.99
GEN	0.16 (0.12)	1.8 (2.39)	43.8 (26.2)	0.028 (0.033)	0.98
GEN	0.19 (0.16)	1.76 (1.2)	54.7 (27.4)	0.029 (0.034)	0.95
GEN	0.22 (0.14)	1.94 (1.36)	68.2 (26.6)	0.04 (0.039)	0.9
GEN	0.26 (0.19)	2.03 (1.52)	72.7 (29.9)	0.046 (0.037)	0.85
GEN	0.32 (0.26)	1.93 (1.4)	75.6 (29.8)	0.056 (0.039)	0.8
GEN	0.31 (0.2)	1.99 (1.47)	79.3 (26.5)	0.06 (0.038)	0.75



(a) EN

(b) GEN

Figure 4: Selected variables for elastic net with and without grouping over 1000 bootstrapping samples of data; each of same size as the trainig set (40 observations). The stopping criteria used were the average values selected using cross validation; (a): $n_z = 40$ non-zero elements for the elastic net, and (b): $n_z = 55$ non-zero elements for the group elastic net with $r_t = 0.95$.

IV. Discussion

The experiments showed an increased generalization power of the group elastic net over the elastic net. The generalization comes from the ability to select groups of highly correlated covariates that are of importance for the prediction at hand. The highly correlated variables are practically averaged over using the L_2-norm shrinkage of the parameter estimates in the model, and hereby noise is averaged out. We have the assumption that it is less likely to have a group of highly correlated variables, which by chance (due to noise) correlates to the output, than it is likely to have a single variable, which by chance correlates to the output. Both looking at estimated prediction errors for test sets and looking at the selection frequency over bootstrapped samples of data confirmed this behavior. As a result, the group elastic net ensured less overfitting and the chance of selecting the irrelevant variables decreased whereas the chance of selecting the relevant variables increased. The increased consistency and the grouping of variables give good interpretation. The consistency makes selected variables trustworthy, i.e. there are fewer falsely detected variables and more correctly detected variables. The grouping gives results, which are easier interpreted for example because of preservation of spatial coherence in images. Such groups of variables would also be an advantage in for example genetic studies where the groups of up or down regulated genes may be found. Taking into account the information from all important covariates seems to be a strong point here. Unfortunately, many of the existing variable selection methods focus on being as sparse as possible, see e.g. (Donoho, 2006), and thereby disregards the information which is present in data with highly correlated covariates. With the proposed method we have one way of more fully exploiting the information present in high-dimensional data sets with naturally mutual correlations amongst covariates. Such data are often seen in image analysis, but also in many other fields, such as for example biostatistics.

V. Conclusion

An algorithm for adaptively grouping highly correlated covariates for prediction was proposed. The algorithm benefits from the grouping ability already present in the elastic net model and simply adds a procedure for selecting groups of covariates into the model using the least angle regression selection algorithm. The algorithm was tested on two data sets and compared to the elastic net model without grouping. The experiments showed an increased consistency in which variables were selected when bootstrapped samples of data were used. Additionally, the group elastic net showed an increased or equivalent generalization power for the test sets. Finally, the groups give enhanced interpretation.

VI. Acknowledgements

The author would like to thank Sara Sharifzadeh, Andrea Mazzaretto and Prof. Allan Aasbjerg Nielsen for fruitful discussions, and Dr. Karl Sjöstrand for providing me with his elastic net Matlab code (Sjöstrand, 2007).

References

Bondell, H., Reich, B. J., 2008. Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar. Biometrics 64, 115–123.

Cho, H., Fryzlewicz, P., 2012. high dimensional variable selection via tilting. Journal of the Royal Statistical Society Series B 74, 593–622.

Clemmensen, L. H., Hansen, M. E., Ersbøll, B. K., 2010. A comparison of dimension reduction methods with applications to multi-spectral images of sand used in concrete. Machine Vision and Applications 21 (6), 959–968.

Clemmensen, L. H., Hastie, T., Witten, D., Ersbøll, B. K., 2011. Sparse discriminant analysis. Technometrics 53 (4), 406–413.

Donoho, D. L., August 2000. High-dimensional data analysis: The curses and blessings of dimensionality. In: Conf. Math Challenges of the 21st Century, Los Angeles.

Donoho, D. L., 2006. For most large underdetermined systems of linear equations the minimal l1-norm solution is also the sparsest solution. Communications on Pure and Applied Mathematics 59 (6), 797–829.

Efron, B., Hastie, T., Johnstore, I., Tibshirani, R., 2004. Least angle regression. Annals of Statistics 32, 407–499.

Hoerl, A. E., Kennard, R. W., 1970. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12, 55–67.

Jin, B., Lorenz, D. A., Schiffler, S., 2009. Elastic-net regularisation: error estimates and active set methods. inverse Problems 25.

Lin, B., Pang, Z., May 2013. Tilted correlation screening learning in thigh dimensional data analysis. Journal of Computational and Graphical Statistics.

Shara, D. B., Bondell, H., Zhang, H. H., 2013. Consistent group identification and variable selection in regression with correlated variables. Journal of Computational and Graphical StatisticsIn Press.

Shen, X., Huang, H. C., 2010. Grouping pursuit through a regularization solution surface. J Am Stat Assoc 105 (490), 727–739.

Sjöstrand, K., 2007. Regularized statistical analysis of atonomy. Ph.D. thesis, Technical Univeristy of Denmark.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of Royal Statistical Society - Series B, 267–288.

Wei, F., Huang, J., 2010. Consistent group selection in high-dimensional linear regression. Bernoulli 16 (4), 1369–1384.

Winham, S., Wang, C., Motsinger-Reif, A. A., 2011. A comparison of multifactor dimensionality reduction and l_1-penalized regression to identify gene-gene interactions in genetic association studies. Statistical Applications in Genetics and Molecular Biology 10 (1).

Yuan, M., Lin, Y., 2006. Model selection and estimation in regression with grouped variables. Journal of Royal Statistical Society - Series B 68 (1), 49–67.

Zou, H., T.Hastie, 2005. Regularization and variable selection via the elastic net. Journal of Royal Statistical Society - Series B 67, 301–320.