Regression and Variable Selection via A Layered Elastic Net

Michael Van B. Supranes and Joseph Ryan G. Lansangan
School of Statistics
University of the Philippines Diliman

One approach in modeling high dimensional data is to apply an elastic net (EN) regularization framework. EN has the good properties of least absolute shrinkage selection operator (LASSO), however, EN tends to keep variables that are strongly correlated to the response, and may result to undesirable grouping effect. The Layered Elastic Net Selection (LENS) is proposed as an alternative framework of utilizing EN such that interrelatedness and groupings of predictors are explicitly considered in the optimization and/or variable selection. Assuming groups are available, LENS applies the EN framework group-wise in a sequential manner. Based on the simulation study, LENS may result to an ideal selection behavior, and may exhibit a more appropriate grouping effect than the usual EN. LENS results to poor prediction accuracy, but applying OLS on the selected variables may yield optimum results. At optimal conditions, the mean squared prediction error of OLS on LENS-selected variables are on par with the mean squared prediction error of OLS on EN-selected variables. Overall, applying OLS on LENS-selected variables makes a better compromise between prediction accuracy and ideal grouping effect.

Keywords: regression, variable selection, variable clustering, high dimensional data, elastic net, grouping effect

1. Introduction

Technological advancements improved retrieval and storage of voluminous data. With the ability of storing more information, many data sets may have too many p variables or n observations, i.e., data sets tend to be high dimensional. High dimensionality imposes many challenges on model building, especially for the case where $n \ll p$. Ordinary least squares (OLS) regression does not have a unique solution when there are more predictors than observations. Having too many predictors also make the model less parsimonious or unlikely interpretable. Existence of multicollinearity is also likely to be observed in a large pool of predictors. These challenges brought by having too many observations or variables are commonly referred to as the “curse of dimensionality”, as coined by Bellman (1957).

One approach in modeling high dimensional data is to reduce the dimension through variable selection or sparse model fitting. Variable selection methods have been developed to effectively reduce the number of variables while targeting an optimum predictive ability. To induce sparsity in modeling and/or to ensure existence of a unique solution, most methods utilize a penalized least squares framework or some regularization in the estimation/optimization method.

Regularization methods such as ridge regression (Hoerl and Kennard, 1970) imposes a penalty $\lambda \geq 0$ on the squared l_2-norm in the estimation of the regression parameters. In effect, the penalty function induces continuous shrinkage of parameters near zero as λ increases (Hastie et al., 2009). There exists a value of λ for which ridge regression has better predictive ability than ordinary least squares, but it does not reduce the regression parameters to nullity. So, ridge regression may not result to an interpretable model in high-dimensional setting where p is much larger than n. Hence, it is commonly used in tandem with best subset selection (or hard thresholding). However, outcomes of hard thresholding are highly varied that small changes in data may lead to a different best subset (Tibshirani, 1996; Brieman, 1996).

To address the weaknesses of ridge regression and best subset selection, Tibshirani (1996) introduced the Least Absolute Shrinkage Selection Operator (LASSO). LASSO uses l_1-norm in the penalty function, which results to sparse regression coefficients (i.e., zero coefficients) unlike that of ridge regression. Moreover, the solution of LASSO coincides with a soft threshold estimator, which results to more precise estimates than that of best subset selection (Tibshirani, 1996). For the case of $n>p$, none of ridge, LASSO and best subset selection globally outperforms the rest, but the sparsity behavior of LASSO becomes preferable as p becomes much larger than n (Zou and Hastie, 2005). However, good properties of LASSO usually assume weak correlations among predictors, which is not safe to consider in high dimensional data. When a group of strongly correlated predictors exists in the data, LASSO tends to select only one variable from the group and does not care which one is selected (Zou and Hastie, 2005).

In this study, a layered selection under the elastic net framework is proposed to mitigate issues on high dimensionality in regression modeling. The procedure is aimed at providing better variable selection (i.e., sparse solution) especially under moderate to high inter-group correlations existing among the predictor variables. The procedure is also intended to have regression coefficient estimates that yield better prediction of the response variable.

In the following section, a brief discussion on improvements of LASSO and other regularization and/or optimization methods under the regression context in the high dimensional setting is presented. In Section 3, the layered elastic net selection or LENS procedure as an alternative is introduced. Simulation parameters and scenarios to evaluate and compare LENS are in Section 4, and results of the simulation study are provided in Section 5. The last section provides concluding remarks.

2. Regularization in High Dimensional Setting

As an improvement to LASSO, Zou and Hastie (2005) introduced the elastic net. The elastic net is a form of regularization that utilizes both penalties in ridge regression and LASSO. The elastic net framework may be formulated as:

$$\boldsymbol{\beta}_{EN} = \text{argmin}_{\beta}\left\{\left\|\boldsymbol{y}-\begin{bmatrix}1 & \boldsymbol{X}\end{bmatrix}\boldsymbol{\beta}\right\|_2^2+\frac{(1-\alpha)}{2}\lambda\left\|\boldsymbol{\beta}\right\|_2^2+\alpha\lambda\left\|\boldsymbol{\beta}\right\|_1\right\} \tag{1}$$

where $\boldsymbol{y}_{nx1}$ is a vector of response variable, $\boldsymbol{X}_{nxp}$ is a matrix of predictors, $\boldsymbol{\beta}_{(p+1)x1}$ is a vector of regression parameters, $\alpha \in [0,1]$ is a mixing parameter, $\lambda \geq 0$ and is a shrinkage parameter. The framework reduces to that of the ridge regression at $\alpha = 0$, while the framework reduces to that of LASSO at $\alpha = 0$ (Friedman et al., 2010).

The introduction of the ridge penalty in the LASSO framework added convexity and flexibility in the sparsity behavior of LASSO. Hence, the elastic net enjoys good properties of LASSO in terms of shrinkage, sparsity and computational costs, while its new-found convexity allows selection of at most p variables even if $n<p$ (Zou and Hastie, 2005). Moreover, the added convexity encourages grouping effect, which means the elastic net selects (or omits) correlated variables in groups (Zou and Hastie, 2005). The grouping effect is strongly observed when the penalty on the ridge component is relatively high or when α is relatively small (Lansangan, 2014).

On the other hand, the grouping effect may not be desirable, when the goal is to identify a very small subset of variables that can maintain a predictive strength comparable to that of a full model (Lansangan, 2014). Predictive power is diminished, when one important feature is dropped out of the model. In addition, it was observed that LASSO and elastic net tend to keep variables that are strongly correlated to the response, which does not necessarily result to the most predictive small subset of predictors. For instance, it is possible for elastic net's grouping effect to retain a relatively unimportant group, when the unimportant group has moderate to strong correlation with relatively more important groups (Lansangan, 2014).

Furthermore, l_p-norm regularized methods provide solutions for several high-dimensional problems, but these methods, even elastic net, do not directly account the grouping structure of variables in model selection. Under the presence of strongly correlations, identifying the grouping structure of variables may result to better selection behavior and/or prediction accuracy. For instance, Sanche and Lonergan (2006) utilized variable grouping together with expert opinion as an initial procedure for strategically selecting predictors. Variable grouping helped in significantly reducing the number of variables to consider (Sanche and Lonergan, 2006). Near non-identifiability of strongly correlated predictors "confuses" variable selection strategies, and the predictive ability of the resulting model could potentially suffer, when variable grouping is neglected (Buhlmann et al., 2013).

One way of accounting correlations among predictors is by constructing indices or principal components. In Linear and Non-replete Selection (LaNS), there is simultaneous dimension reduction and variable selection, which resulted to construction of sparse principal components that are optimal for predicting the response. Based on simulation studies, LaNS may perform at par or better than LASSO and elastic net in terms of selection behavior and prediction accuracy (Lansangan and Barrios, 2017).

In other methods, predictors must be classified into groups before model fitting and/or variable selection. Grouped versions of existing sparsity-inducing methods (e.g. LASSO, Least Angle Regression, Non-negative Garrotte) were developed to address the problem of handling strong correlations in variable selection (Yuan and Lin, 2006). Group LASSO imposes different penalty sizes across predefined groups of variables, and is found to perform at par or better than LASSO and LARS prediction-wise (Yuan and Lin, 2006). Friedman et al. (2010) developed Sparse Group LASSO, which may result to sparse coefficients for each group. Sparse Group LASSO exhibits a good compromise between Group LASSO and LASSO, yielding grouped sparsity in the fitted model (Friedman et al., 2010). In a study by Buhlmann et al. (2013), hierarchical agglomerative grouping algorithms were utilized in developing clustered LASSO methods. Implemented using Canonical Correlations, Clustered Representative LASSO (CRL) and Clustered Group LASSO (CGL) were shown to be effective variable screening methods, as it effectively keeps all groups with at least one significant predictor (Buhlmann et al., 2013).

Studies conducted by Lansangan and Barrios (2017), Buhlmann et al. (2013), and Friedman et al. (2010) show that grouping variables, whether by constructing indices or clustering, may help in attaining strong predictive ability and desirable selection behavior. This study provides an alternative framework of utilizing the elastic net such that interrelatedness and groupings of predictors are explicitly considered in variable selection. The idea is to select within groups of important predictors, instead of selecting among groups (i.e. the grouping effect). Details of the method, and comparison of alternative and existing elastic net procedures are discussed in the following sections.

3. Layered Elastic Net Selection

Suppose that strongly correlated predictors were grouped together prior to modeling. Then, a preferred selection strategy may be applied locally in each group of variables, such that "sufficient representation" is achieved. An elastic net regularization framework applied on a group of predictors in a sequential manner is proposed, such procedure and/or regularization will be referred to as the Layered Elastic Net Selection (LENS). The LENS procedure requires predetermined grouping of the predictor variables prior to the variable selection optimization. The groups are assumed to be non-overlapping and are those that may (linearly) predict the response variable.

Suppose grouping of the predictor variables is available. Let G be the number of groups. LENS may then be applied for variable selection and model fitting. Let $\boldsymbol{y}_{nx1}$ be a vector of response variable, $\mathbf{X}_{nxp}$ be a matrix of predictors, and $[\boldsymbol{X}_1 \quad \boldsymbol{X}_2 \ldots \boldsymbol{X}_k \ldots \boldsymbol{X}_G]$ is a partition of $\boldsymbol{X}$, where $\boldsymbol{X}_k$ is the k^{th} group of predictors. The LENS is specified below.

Layered Elastic Net Selection (LENS) Algorithm

1. Identify groups of strongly correlated predictors through a grouping algorithm.
2. For each group $\boldsymbol{X}_k$, derive a synthetic $\boldsymbol{z}_k = \boldsymbol{X}_k\boldsymbol{v}_{1,k}$, where $v_{1,k}$ is the first column of $\boldsymbol{V}_k$ and $\mathrm{U}_k\mathrm{D}_k\mathrm{V}_k^T$ is the singular value decomposition of $\boldsymbol{X}_k$. Let $\boldsymbol{Z} = [\boldsymbol{z}_1 \quad \boldsymbol{z}_2 \ldots \boldsymbol{z}_k \ldots \boldsymbol{z}_G]$.
3. For each group $\boldsymbol{X}_k$, compute an impact score $s_k = \frac{1}{n}\left\|\boldsymbol{y} - \boldsymbol{Z}_{-k}\widehat{\boldsymbol{\beta}_k}\right\|_2^2$ where $\widehat{\boldsymbol{\beta}_k}$ is the OLS estimate when $\boldsymbol{y}$ is regressed against $\boldsymbol{Z}_{-k}$ (i.e. matrix $\boldsymbol{Z}$ without the k^{th} column).
4. Variable groups are sequenced in decreasing order with respect to s_k. Let $\boldsymbol{X} = [\boldsymbol{X}_{(1)} \quad \boldsymbol{X}_{(2)} \ldots \boldsymbol{X}_{(k)} \ldots \boldsymbol{X}_{(G)}]$ where $\boldsymbol{X}_{(k)}$ is the k^{th} group in the sequence.
5. For a given γ_1 and γ_2, apply Elastic Net on each group in a sequential manner:

   ```
   DO for $k = 1,2,\ldots,G$.
     SOLVE for $\widehat{\boldsymbol{\beta}_k}$ as:
       $\hat{\boldsymbol{\beta}}_k = \arg\min_{\boldsymbol{\beta}_k}\left\{\left\|\boldsymbol{y}_k - [1 \quad \boldsymbol{X}]\boldsymbol{\beta}_k\right\|_2^2 + \gamma_2\left\|\boldsymbol{\beta}_k\right\|_2^2 + \gamma_1\left\|\boldsymbol{\beta}_k\right\|_1\right\}$
     IF $k = 1$, THEN $\boldsymbol{y}_k = \boldsymbol{y}$. ELSE, $\boldsymbol{y}_k = \boldsymbol{y}_{k-1} - \boldsymbol{X}_{k-1}\hat{\boldsymbol{\beta}}_{k-1}$.
   END of DO.
   ```

6. The intercept $\hat{\boldsymbol{\beta}}_0$ is computed as mean of $\boldsymbol{y}_G - \boldsymbol{X}_G\hat{\boldsymbol{\beta}}_G$.
7. The LENS estimate of coefficients is

   $$\hat{\boldsymbol{\beta}}_{LENS}^T = \left[\hat{\boldsymbol{\beta}}_1^T \quad \hat{\boldsymbol{\beta}}_2^T \cdots \quad \hat{\boldsymbol{\beta}}_k^T \cdots \quad \hat{\boldsymbol{\beta}}_g^T \quad \hat{\boldsymbol{\beta}}_0\right].$$

In case there is no known or intuitive grouping of variables, a data-driven variable grouping algorithm is also proposed. In this study, the variable grouping procedure is based on a specific hierarchical agglomerative grouping (HAG) algorithm. As in any HAG procedure, the choice of dissimilarity measure and linkage method is important. Since the goal is to group strongly correlated predictors, the dissimilarity measure between the ith variable and the jth variable will be defined as $d(i, j) = 1-|r_{ij}|$, where r_{ij} is the Pearson's correlation between the two variables. The value of d(i,j) is small when two variables are strongly correlated or similar. The same dissimilarity measure was considered in Buhlmann et al. (2013). For group linkage, average linkage will be used. It is preferred over single linkage and complete linkage, because average linkage is less likely to exhibit crowding or chaining (Hastie et al., 2009).

Let $\boldsymbol{X}$ be an nxp matrix of predictors, then the hierarchical agglomerative grouping algorithm is as follows:

Hierarchical Agglomerative Grouping of Variables (HAG Var)

```
INITIALIZE $G=p$.
DO WHILE $G>1$. (Hierarchical Grouping)
    PARTITION the data set into G groups. (For $G=p$, each group contains
        one variable) $\boldsymbol{X} = [\boldsymbol{X}_1 \ldots \boldsymbol{X}_l \ldots \boldsymbol{X}_{(G)}]$ ,where $\boldsymbol{X}_1$ is a $nxp_1$ matrix and
        $\sum p_1 = p$
    FOR all $m < n$, CALCULATE the dissimilarity $d(m,n)$ bet. groups $X_m$
        and $X_n$ as:
            $d(m,n) = \frac{1}{p_n p_m} \sum_{j=1}^{p_n} \sum_{i=1}^{p_m} \left(1 - \left|r_{m(i),n(j)}\right|\right)$
            ($average\ dissimilarity\ bet.groups\ m\ and\ n$)

    Where $r_{m(i),n(j)} = cor(\boldsymbol{x}_{m(i)}, x_{n(j)})$ is the Pearson's correlation of $i^{th}$ variable in
        group $cor(\boldsymbol{x}_{m(i)})$ and $j^{th}$ variable in group $n(x_{n(j)})$.
    END FOR.
    UPDATE $G = G$-1 by taking $\boldsymbol{X}_m$ and $\boldsymbol{X}_n$ with minimum $d(m,n)$ as one
        group.
END DO.
OUTPUT Dendogram.
```

The grouping algorithm results to a dendogram. The next step is to identify a suitable number of groups. The minimum group average silhouette index (or MiGASi index) will be used as an internal validation measure. For a given a grouping solution, the average silhouette is measured per group. MiGASi is the smallest computed group average silhouette index. The silhouette index measures the appropriateness of a member-variable with respect to its current group assignment (Rousseeuw, 1987). Let $a(i)$ be the average dissimilarity of object i to all other members of the group where i belongs. Let $b(i)$ be the lowest average dissimilarity of object i to any other group, of which i is not a member. The silhouette of i, $s(i)$, is defined as:

$$s(i)=\frac{b(i)-a(i)}{\max\left(a(i),b(i)\right)}$$

The silhouette is bounded by -1 and 1, where a value close to 1 is interpreted as being appropriately assigned to its current group and a value close to -1 means that the member should be assigned to one of its neighboring groups. Hence, the MiGASI index measures the tightness of groups vis-à-vis separability of groups. If MiGASi index is closed to 1, then it means that most (if not all) variables are appropriately grouped and groups are tight and homogeneous. For simplicity, MiGASi will be measured for the last 20 nodes or joints of the dendogram. These values serve as the candidate subset for determining the right number of groups. The dendrogram will be cut at the node yielding the maximum MiGASi, which will result to G number of variable groups.

4. Design of Simulation Study

The predictive ability and selection behavior of elastic net and LENS are compared through a simulation study. The data generating process involves three hidden factors that will generate three groups of strongly correlated predictors. The simulation strategy is similar to that of Zou and Hastie (2005), and Lansangan and Barrios (2017). The hidden factors f_1, f_2 and f_3 represent the "key drivers" or important features that predict the response variable. The hidden factors were generated such that:

$$f_1 \sim \mathrm{N}\ (400,\ 150^2), f_2 \sim \mathrm{N}\ (400,\ 145^2), f_3 \sim N\ (400,\ 125^2)$$

$$cor\left(f_1,f_2,f_3\right)=\begin{bmatrix} 1 & 0 & 0.6 \\ 0 & 1 & -0.5 \\ 0.6 & -0.5 & 1 \end{bmatrix}$$

Given those specifications, f_1 and f_2 are independent of each other, while f_3 is moderately correlated to other factors. This represents that case where there is one important data feature which may be confused with the other two.

In simulating a data set, p independent variables were generated as a function of a hidden factor (i.e. either $f_1 f_2$ or f_3). Three groups of variables were simulated, as follows:

$$X_{ij} = f_1 + N\left(0, \left(1 + \frac{j}{25}\right)15\right), j = 1,2\ldots15$$

$$X_{ij} = f_2 + N\left(0, \left(1 + \frac{j-25}{25}\right)15\right), j = 16,15\ldots30$$

$$X_{ij} = f_3 + N\left(0, \left(1 + \frac{j}{500}\right)10\right), j = 31,32\ldots p$$

Each X_{ij} represents a measurable manifestation of a hidden factor. Since $X_{ij}s$ are linear function of a hidden factor, variables generated from the same hidden factor are strongly correlated with each other. In addition, predictors from the first two hidden factors are moderately correlated to predictors from f_3.

The response variable is generated as a linear combination of predictors and an error term. The model is:

$$y_i \mid \boldsymbol{x}_i^T = X_{i1}\beta_1 + X_{i2}\beta_2 + \ldots + X_{ip}\beta_{1000} + \varepsilon_i$$

$$\textit{where } \varepsilon_i \sim N(0,50^2)\ \forall i, \textit{ and } cov\ (\varepsilon_i, \varepsilon_k) = 0\, \forall i \neq k.$$

The β_js were selected such that group 1 has the largest relative contribution to the value of response, group 2 has the second largest contribution, and group 3 has the least contribution.

Two different scenarios were considered in the simulation study—a non-high dimensional case (NHD), and a high dimensional (HD) case. Each scenario was replicated 200 times. One hundred predictors were generated for NHD, and 1000 predictors were generated for HD. Table 1 shows the simulated relative contribution of each group for each scenario.

Table 1. Scenario Settings for Simulation Study

Case	Relative Contribution by Group		
	Group 1	Group 2	Group 3
NHD	~48%	~35%	~15%
HD	~47%	~34%	~18%

For both scenarios, the ideal result is to select most variables from group 1 and group 2, since ~80% of the value of response variable comes from these groups. Nevertheless, group 3 still has a significant contribution to the response variable. Hence, it is still expected to get optimal results when there are few variables selected from group 3. In addition, the ideal result for the grouping effect is to keep variables from group 1 when very few variables are taken into consideration.

In the simulation design, the contribution of the smaller groups of variables (i.e., group 1 and group 2) are larger than group 3 to mimic the case where few variables (i.e. 30 out of 100, or 30 out of 1000) are sufficient to explain most variation in the response. Hence, it makes sense to select and concentrate on very few predictors in constructing the model, and thus key to a regression solution is to find an "optimal small subset" of predictors.

Several values of tuning parameters are utilized in comparing elastic net and LENS in terms of the behavior of selection and grouping effect. Table 2 shows all methods and settings for comparison.

Table 2. Methods in Comparison

Elastic Net **EN(α, λ)**	**Layered Elastic Net Selection** **LENS(γ_1, γ_2)**
EN(0.999, λ^*)	LENS(γ^*, 0.1)
EN(0.900, λ^*)	LENS(γ^*, 50)
OLS-EN(0.999, λ^*)	LENS(γ^*, 100)
OLS-EN(0.900, λ^*)	LENS(γ^*, 300)
	LENS(Best)
	OLS-LENS(γ^*,0.1)
	OLS-LENS(γ^*,50)
	OLS-LENS(γ^*,100)
	OLS-LENS(γ^*,300)
	OLS-LENS(Best)
	Best OLS-LENS

For Elastic Net (EN), two values of mixing parameter α are considered (i.e. 0.999 and 0.900). When $\alpha = 0.999$, the elastic net framework is almost the same as the LASSO framework. When $\alpha = 0.900$, the penalty on the l_2-norm of parameters is relatively higher. Hence, the second setting is somehow more "elastic" and the grouping effect is expected to be observed (Lansangan, 2014). For LENS, varying values of ridge penalty γ_2 were considered. When $\gamma_2 = 0.1$, each stage in layered selection closely resembles a LASSO. On the other hand, higher values of γ_2 were considered to explore the grouping effect at increasing ridge penalty. For both methods, λ^* and $\gamma^* = \gamma_1^*$ are values of the tuning parameter which retain a target number of predictors or less for fixed α and γ_2, respectively. This was operationalized by setting the target number of variables to 15.

In addition to elastic net and LENS, the ordinary least squares (OLS) counterpart of each model is also considered (denoted as OLS-EN or OLS-LENS). OLS is applied on the selected predictors after conducting EN or LENS. Since very few predictors will be selected, OLS will have a unique solution even under the high dimensional case.

The best LENS specifications, denoted as LENS (*Best*), OLS-LENS (*Best*) and *Best* OLS-LENS, are identified for each iteration. LENS (*Best*), is the LENS specification which attains the lowest BIC score. OLS-LENS is the OLS counterpart of LENS (*Best*), (i.e., regression estimates based on the OLS fit for the selected variables). While for *Best* OLS-LENS, the BIC-type criterion was computed and evaluated after fitting the OLS counterpart of candidate LENS specifications. The *Best* OLS-LENS is then the OLS-LENS with the lowest BIC.

In comparing the methods, the Mean Squared Prediction Error (MSPE) is computed to evaluate the predictive ability of the model. Let $\boldsymbol{y}$ be the response vector and $\hat{\boldsymbol{y}}$ be the vector of predicted values. The MSPE is defined as $MSPE = \frac{1}{n}\left\|\boldsymbol{y}-\hat{\boldsymbol{y}}\right\|_2^2$. The BIC-type criterion by Zou et al. (2007) was computed to evaluate the prediction accuracy vis-à-vis variable selection (or the number of predictors in the model). Let *n* be the number of observations and *NNZ* be the number of nonzero coefficients in the model. The BIC-type criterion is defined as:

$$BIC^* = \frac{MSPE}{var(\boldsymbol{y})} + NNZ\frac{\log(n)}{n}$$

The BIC-type criterion penalizes the measure of prediction accuracy with the number of nonzero coefficients and the number of observations. Hence, the model with the lowest BIC-type criterion tends to be the most parsimonious, i.e. the model has the smallest prediction error at a small number of predictors possible. The average number of selected predictors by group is also computed for describing the selection behavior of each method.

5. Results and Discussion

The results are organized into 3 sections. The first section covers simulation results under the NHD case, while the second section covers the HD case. Discussions on the choice of number of groups prior to use of LENS are in the last section.

Non-high dimensional case

For the non-high dimensional case, there are 200 observations and 100 predictors. Seventy out of 100 predictors are from group 3, which is correlated to the more important groups (i.e. groups 1 and 2). For selection behavior, the ideal outcome is then to get more variables from group 1, followed by group 2, and fewest from group 3. Table 3 summarizes the results for the non-high dimensional case.

Table 3. Results for Non-High Dimensional Case

Methods	Average Number of Variables Selected from			Average MSPE	Average BIC
	Group 1	Group 2	Group 3		
EN(0.999, λ^*)	5.84	5.18	3.91	3 167 157	0.422
EN(0.900, λ^*)	0.66	3.23	11.11	481 496 979	4.442
LENS(γ^*,0.1)	7.36	4.59	2.99	42 451 572	0.752
LENS(γ^*,50)	8.04	4.06	2.85	43 862 207	0.764
LENS(γ^*,100)	8.85	3.50	2.61	45 261 854	0.777
LENS(γ^*,300)	10.96	3.34	0.70	49 903 853	0.818
LENS(*Best*)	8.40	4.59	1.97	42 320 575	0.749
OLS-EN(0.999, λ^*)	5.84	5.18	3.91	[1]267 120	0.398
OLS-EN(0.900, λ^*)	0.66	3.23	11.11	31 816 886	0.664
OLS-LENS(γ^*,0.1)	7.36	4.59	2.99	[1]307 127	0.398
OLS-LENS(γ^*,50)	8.04	4.06	2.85	319 707	0.399
OLS-LENS(γ^*,100)	8.85	3.50	2.61	347 237	0.399
OLS-LENS(γ^*,300)	10.96	3.34	0.70	1 075 452	0.406
OLS-LENS(*Best*)	8.40	4.59	1.97	397 446	0.397
Best OLS-LENS	8.25	4.50	2.20	[1]283 466	0.396

NOTE: [1]Three lowest MSPE.

Both EN and LENS showed ideal results for specifications where α is closed to 1 and γ_2 is closed to 0, respectively. On the other hand, the allocation of EN(0.999, λ^*) tends to be almost equal among groups, while LENS(γ^*, 0.1) tends to select half of predictors from group 1 (which attributes to ~50% of the value of the response). For the grouping effect, the ideal outcome is to keep variables from group 1, since group 1 has the most relative contribution to the value of the response. Consistent with the observation of Lansangan (2014), EN exhibits the grouping effect, when the penalty on the ridge component is relatively high (i.e. α = 0.900 in the simulation study). However, EN(0.900, λ^*) tends to retain variables from group 3, which is the least important group in terms of prediction. LENS, in contrast, tends to retain group 1 in the model at very high values of γ_2 (i.e. LENS(λ^*,300)). Comparing LENS and EN, simulation results suggest that the LENS procedure may arrive at better results than EN with respect to the grouping effect.

Table 3 also summarizes the MSPE and BIC of the models. In terms of prediction error and BIC, EN(0.999, λ^*), i.e. the better specification for EN, outperforms LENS(*Best*). On the other hand, LENS outperforms EN in terms of prediction error and BIC at specifications where the grouping effect is very evident. From Table 3, the average BIC of EN(0.900,λ^*) is 4.442 and the average BIC of LENS(γ^*, 300) is 0.818. This is because EN(0.900,λ*) tend to keep the least important group while LENS(γ^*, 300) tend to keep the most important group.

For both methods, conducting OLS on selected variables greatly improves the MSPE and BIC. The improvement in MSPE is very evident for LENS. For instance, the average MSPE of LENS(*Best*) is 42,320,575, while the average MSPE of OLS-LENS(*Best*) is as low as 397,446. Moreover, it seems that identifying the best among OLS-LENS (i.e. *Best* OLS-LENS) results to better prediction error than taking the OLS counterpart of LENS(Best) (i.e. OLS-LENS(*Best*)). The MSPE of *Best* OLS-LENS is 283,466; which is better than OLS-LENS(*Best*) and on a par with OLS-EN(0.900, λ^*) with average MSPE = 267,120. In addition, OLS-EN(0.900, λ^*), OLS-LENS(*Best*), and *Best* OLS-LENS have comparable average BIC scores.

Table 4. Results for High Dimensional Case

Methods	Average Number of Variables Selected from			Average MSPE	Average BIC
	Group 1	Group 2	Group 3		
EN(0.999, λ^*)	5.62	4.66	4.64	3 867 767	0.426
EN(0.900,λ^*)	0.06	0.20	14.67	1 117 415 971	9.478
LENS(γ^*,0.1)	8.15	4.71	2.12	46 700 940	0.773
LENS(γ^*,50)	8.67	3.96	2.37	48 254 038	0.787
LENS(γ^*,100)	9.33	3.38	2.28	50 388 384	0.805
LENS(γ^*,300)	11.42	3.12	0.46	54 729 130	0.842
LENS(*Best*)	8.40	4.59	1.97	46 259 940	0.769
OLS-EN(0.999,λ^*)	5.62	4.66	4.64	[1]294 568	0.397
OLS-EN(0.900,λ^*)	0.06	0.20	14.67	87 131 031	1.097
OLS-LENS(γ^*,0.1)	8.15	4.71	2.12	[1]335 737	0.399
OLS-LENS(γ^*,50)	8.67	3.96	2.37	348 931	0.400
OLS-LENS(γ^*,100)	9.33	3.38	2.28	477 531	0.401
OLS-LENS(γ^*,300)	11.42	3.12	0.46	1 847 072	0.412
OLS-LENS(*Best*)	8.40	4.59	1.97	528 029	0.400
Best OLS-LENS	8.25	4.50	2.20	[1]320 701	0.398

NOTE: [1]Three lowest MSPE.

High dimensional case

For the HD case, there are 200 observations and 1000 predictors. Nine hundred seventy out of 1000 predictors are from group 3, which is correlated to more important groups (i.e. groups 1 and 2). As in the NHD case, for the selection behavior, the ideal outcome is to get more variables from group 1, followed by group 2, and fewest from group 3. Table 4 summarizes the results for the high dimensional case.

Both EN and LENS showed ideal results for specifications where α is closed to 1 and γ_2 is closed to 0, respectively. The allocation of EN(0.999, λ^*) also tends to be almost equal among groups, while LENS(γ^*, 0.1) tends to select half of predictors from group 1. Consistently, EN(0.900, λ^*) and LENS(γ^*, 300) exhibits the grouping effect, as it tends to keep one group of variables only. For the HD case, the LENS framework showed a better behaving grouping effect than EN, as EN tends to keep the least important group while LENS tend to keep the most important group.

In terms of prediction error and BIC, EN(0.999, λ^*), i.e. the better specification for EN, outperforms LENS(*Best*), i.e. the best specification for LENS. On the other hand, LENS outperforms EN in terms of prediction error and BIC at specifications where the grouping effect is very evident. From Table 4, the average BIC of EN(0.900, λ^*) is much greater than that of LENS(γ^*, 300), which is most likely because of the difference in their grouping effect.

Like the NHD case, conducting OLS on selected variables greatly improves the MSPE and BIC. The improvement in MSPE is very evident for LENS. Consistently, identifying the best among OLS-LENS (i.e. *Best* OLS-LENS) results to better prediction error than taking the OLS counterpart of LENS(*Best*) (i.e. OLS-LENS(*Best*)). The MSPE of *Best* OLS-LENS is 320,701; which is much better than OLS-LENS(*Best*). It seems that the difference between OLS-LENS(*Best*) and *Best* OLS-LENS is more evident in the high dimensional case. Moreover, this adds to the evidence that LENS tends to select an optimal subset of predictors, but its fitted model results to poor prediction. In general, LENS is best treated as a selection operator, and applying OLS on the selected variables may yield optimum results.

The lowest average MSPE (i.e. 294,568) is also attained by OLS-EN(0.999, λ^*). In the HD case, applying OLS after EN seems to be slightly better than applying OLS after LENS. However, applying OLS on LENS-selected variables makes a good compromise between prediction accuracy and ideal grouping effect, because EN seems to be more sensitive to relationships across groups of predictors. In addition, simulation study shows that OLS-EN(0.900, λ^*), and Best OLS-LENS have comparable average BIC scores. OLS-EN and OLS-LENS tend to have similar behavior with respect to minimizing the prediction error at a small number of predictors.

Simulation results for hierarchical agglomerative grouping

As LENS requires predictors to be grouped, for this study, a hierarchical agglomerative grouping algorithm was implemented where the distance measure is based on pairwise-correlation. In addition, selection of the optimal number of groups was automated using the MiGaSi index. The number of groups is chosen such that MiGaSi index is maximized.

Based on the simulation, the outcome of the grouping algorithm affects the result of LENS. The MiGaSi index does not always result to the right number of groups (G), which is 3 in this case. The MiGaSi index suggested $G=3$ about 50% of the time, while it suggested $G=2$ for other cases. As presented in Table 5, it was observed that optimum results for LENS were attained when OLS was applied on the selected variables and the optimal number of groups was identified. For both cases, the average MSPE of Best OLS-LENS is better for cases where the suggested value of G is 3. Under optimal conditions, OLS-LENS and OLS-EN results to comparable prediction accuracy.

Table 5. Average MSPE of OLS-EN and OLS-LENS at Optimum Settings

Case	Method	Average MSPE	Number of Replicates
NHD	OLS-EN(0.999, λ^*)	267 120	200
	Best OLS-LENS (All replicates)	283 466	200
	Best OLS-LENS (Replicates with $G=3$)	276 522	110
HD	OLS-EN(0.999, λ^*)	294 568	200
	Best OLS-LENS (All replicates)	320 701	200
	Best OLS-LENS (Replicates with $G=3$)	305 653	106

Figure 1 shows the average MiGaSi index for different values of G. On the average, the MiGaSi index peaks at the correct number of groups ($G=3$), which makes the criterion a reasonable basis for variable grouping. For cases where *G* was suggested, the average MiGaSi clearly peaks at $G=3$. In cases where $G=2$ was suggested, the MiGaSi index of $G=2$ was slightly higher than that of $G=3$. The pattern was consistently observed in all simulated scenarios. Thus, it is recommended to consider all values of G with near optimal MiGaSi index, most especially if the MiGaSi index values are very near to each other. This may assure that the most appropriate value of G was considered in model selection.

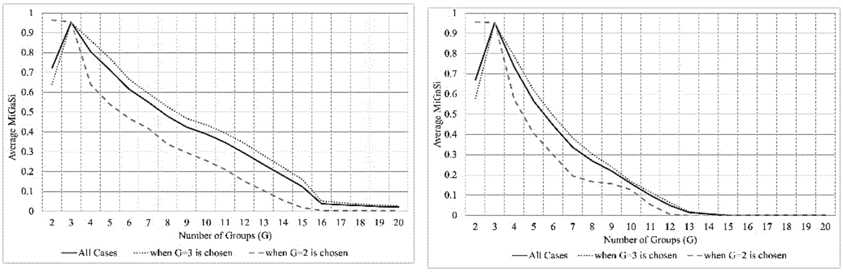

Figure 1: Average MiGaSi index for each Number of Groups in NHD (left) and HD (right)

6. Conclusions

Layered Elastic Net Selection (LENS) applies the elastic net regularization framework in groups of variables in a sequential manner. Based on the simulation study, LENS is a good selection operator and may result to a better selection behavior than the elastic net. In the case that there are groups of strongly correlated predictors (i.e. important hidden features) but only a small subset of important predictors exists, the LENS may be able to select few variables that capture all important features, and the more important groups are more likely prioritized. In addition, LENS may result to a more appropriate grouping effect. Under the assumption of moderate inter-group dependencies, the grouping effect of elastic net is expected to select the "weaker" group. LENS, in contrast, given optimal values of G, γ_2, γ_1, may minimize the chance of keeping relatively "weaker" groups, and thus may select only the most important groups.

The estimated coefficients of LENS $\hat{\beta}_{LENS}$ however results to poor prediction accuracy, but the applying OLS on the selected variables results to an improved predictive ability. Hence, LENS makes a good selection operator, and it is most appropriate to apply OLS on selected variables in fitting the predictive model (similar to LaNS, cf. Lansangan and Barrios, 2017).

In general, the number of groups (G), value of tuning parameters (γ_1, γ_2), and the sequence of fitting are important considerations in applying LENS. Attaining optimal results are dependent on the chosen values. Based on the simulation studies, there is a value of G, γ_1, and γ_2 where OLS on LENS-selected variables results to a model that is on par with elastic net in terms of prediction accuracy vis-à-vis parsimony. For the selection of tuning parameters, setting a target number of variables may help in operationalizing the search for γ_1 and γ_2.

Another important consideration in LENS is the sequence of fitting. The sequential approach of LENS differentiates it from existing strategies. Unlike other grouped sparse method, the LENS framework utilizes much fewer tuning parameters as existing methods tend to require separate $\lambda_{1,k}$ for each group $\boldsymbol{X}_k$

and the search for optimal tuning parameters is relatively more tedious and complicated. In lieu of varying tuning parameters, the framework requires a ranking algorithm. The LENS algorithm may be viewed as a modified backfitting algorithm, i.e., the sequence of fitting is crucial in LENS as in backfitting.

REFERENCES

BELLMAN, E., 1957, *Dynamic Programming*, (Rand Corporation) Princeton University Press.

BREIMAN, L., 1996, Heuristics of instability and stabilization in model selection, *The Annals of Statistics* 24(6):2350-2383.

BUHLMANN, P., and VAN DE GEER, S., 2011, *Statistics for High-Dimensional Data, Springer Series in Statistics*, Springer Heidelberg Dordrecht London New York.

BUHLMANN, P., RUTIMANN, P., VAN DE GEER, S., and ZHANG, C., 2013, Correlated Variables in Regression: Clustering and Sparse Estimation, *Journal of Statistical Planning and Inference* 143:1835-1858.

CHAVENT, M., KUENTZ, V., LIQUET, B., and SARACCO, J., 2011, ClustofVar: An R Package for the Clustering of Variables, *Journal of Statistical Software*. [Downloaded September 2015: http://arxiv.org/pdf/1112.0295.pdf]

DONOHO, D., 2000, High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality. [Downloaded September 2015: http://signallake.com/innovation/Curses.pdf]

FRIEDMAN, J., HASTIE, T., and TIBSHIRANI, R., 2010, A Note on the Group Lasso and a Sparse Group Lasso. [Downloaded September 2015: http://statweb.stanford.edu/~tibs/ftp/sparse-grlasso.pdf]

FRIEDMAN, J., HASTIE, T., and TIBSHIRANI, R., 2010, Regularization Paths for Generalized Linear Models via Coordinate Descent, Journal of Statistical Software. 33(1), 1-22. [Downloaded September 2015: http://www.jstatsoft.org/v33/i01/]

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J., 2009, *The Elements of Statistical Learning* 2nd Ed., Springer.

HOERL, A., and KENNARD, R., 1970, Ridge Regression: Applications to Nonorthogonal Problems, *Technometrics* 12(1):69-82.

JOHNSON, R., and WICHERN, D., 1988, *Applied Multivariate Statistical Analysis*, Prentice-Hall, Inc. Upper Saddle River, NJ, USA.

LANSANGAN, J.R., 2014, Simultaneous Dimension Reduction and Variable Selection in Modeling High Dimensional Data (Unpublished doctoral dissertation), School of Statistics, University of the Philippines.

LANSANGAN, J.R.G and BARRIOS, E.B., 2017, Simultaneous Dimension Reduction and Variable Selection in Modeling High Dimensional Data, *Computational Statistics and Data Analysis* Vol. 112 (August 2017), p. 242-256.

LEE., T., DULING, D., SONG, L., and LATOUR, D., 2008, SAS Global Forum Paper: Two-Stage Variable Clustering for Large Data Sets. [Downloaded September 2015: http://support.sas.com/resources/papers/sgf2008/2stagecluster.pdf]

MAECHLER, M., ROUSSEEUW, P., STRUYF, A., HUBERT, M., and HORNIK, K., 2015, *Cluster: Cluster Analysis Basics and Extensions*, R package version 2.0.3. [Downloaded September 2014 20: https://cran.r-project.org/web/packages/cluster/index.html]

R CORE TEAM, 2015, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. [URL: http://www.R-project.org]

ROUSSEEUW, P., 1987, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, *Journal of Computational and Applied Mathematics*, 20:53-65.

SANCHE, R., and LONERGAN, K., 2006, Casualty Actuarial Society Forum Paper: Variable Reduction for Predictive Modeling with Clustering. [Downloaded September 2015: http://www.casact.org/pubs/forum/06wforum/06w93.pdf]

SEARLE, S., 1982, *Matrix Algebra Useful for Statistics, Wiley Series in Probability and Mathematical Statistics*, John and Sons, Inc., Hoboken, New Jersey.

TIBSHIRANI, R., 1996, Regression Shrinkage and Selection via the Lasso, *Journal of the Royal Statistical Society*, Series B (Methodological). 58(1):267-288

YUAN, M., and 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. and HASTIE, T., 2005, Regularization and Variable Selection via the Elastic Net, *Journal of Royal Statistical Society* Series B 67(2):301-320.

ZOU, H., and HASTIE, T., 2006, Elasticnet: Elastic-Net for Sparse Estimation and Sparse PCA. R package version 1.1. [Downloaded September 2014: http://CRAN.R-project.org/package=elasticnet]

ZOU, H., HASTIE, T., and TIBSHIRANI, R., 2007, On the “Degrees of Freedom” of the LASSO, *The Annals of Statistics* 35(5): 2173- 2192.