A Novel Two-step Sparse Learning Approach for Variable Selection and Optimal Predictive Modeling

Yiren Liu [*,**], **S. Joe Qin** [*]

[*] *School of Data Science and Hong Kong Institute for Data Science*
City University of Hong Kong, 83 Tat Chee Ave., Hong Kong
E-mail: joe.qin@cityu.edu.hk; Corresponding author.
[**] *Centre for Systems Informatics Engineering*
Shenzhen Research Institute of CityU, Shenzhen, China

Abstract:
In this paper, a two-step sparse learning approach is proposed for variable selection and model parameter estimation with optimally tuned hyperparameters in each step. In Step one, a sparse learning algorithm is applied on all data to produce a sequence of candidate subsets of selected variables by varying the hyperparameter value. In Step two, for each subset of the selected variables from Step one, Lasso, ridge regression, elastic-net, or adaptive Lasso is employed to find the optimal hyperparameters with the best cross-validation error. Among all subsets, the one with the overall minimum cross-validation error is selected as globally optimal. The effectiveness of the proposed approach is demonstrated using an industrial NOx emission dataset and the Dow challenge dataset to predict product impurity.

Keywords: Inferential modeling; sparse statistical learning; variable selection; regularization; industrial applications

1. INTRODUCTION

Inferential sensors have been studied and practiced in process industries for over three decades to predict hard-to-measure quality variables from easy-to-measure process variables (Tham et al. (1991); Qin and McAvoy (1992); Qin et al. (1997); Galicia et al. (2011); Khatibisepehr et al. (2013); Shang et al. (2014)). Not only are inferential sensors useful for product quality prediction and monitoring, they are also integrated into model predictive control systems to provide feedback control (Zhao (2021); Kano et al. (1998)). Zhao (2021) reported that Aspen Technology has deployed over 16,000 inferential sensors in the manufacturing plants of their clients world-wide.

One of the critical tasks in developing data-driven inferential sensors is to select relevant variables to build the best predictive model. In practice, variable selection for inferential sensors has largely relied on experience or trial-and-errors. Recently, sparse statistical learning methods such as the least absolute shrinkage and selection operator (Lasso) family of algorithms have provided promising solutions to select relevant variables (Hastie et al. (2015)). Lasso introduces an l_1 norm penalty on the magnitude of the model coefficients with a tuning hyperparameter to suppress the coefficients of irrelevant variables. The Lasso uses cross-validation (CV) to select an optimal hyperparameter for the best generalization. After the optimal hyperparameter is determined, a final model is re-trained on all training data. The sparse methods introduce a bias in the model, but they often outperform unbiased models by achieving a favorable reduction in the variance.

Although the Lasso can lead to zero coefficients for some variables, they suffer from several potential drawbacks when applied to process data, which are usually collinear due to material and energy balances. One of the problems is that the selected set of variables can be sensitive to the data samples used in determining the hyperparameter. Minor changes in the training data can change the selected variables from one subset to another without a sensible improvement in the model quality. This is more pronounced for process variables with collinearity. This problem has been reported in Kamkar et al. (2015); Arora and Kaur (2020); Meinshausen and Bühlmann (2010); Sun et al. (2013); Qin and Liu (2021), where stable Lasso solutions have been proposed.

Another drawback of Lasso is its acclaimed advantage, in that Lasso finds zero coefficients and non-zero coefficients in one step with one hyperparameter value. It is clear that Lasso can drive the coefficients of some variables to zero by tuning the hyperparameter, but the same hyperparameter value does not necessarily yield optimal estimates for the non-zero coefficients of the remaining variables. This issue is pointed out in Meinshausen (2007), where a relaxed solution was proposed.

In this paper, we propose a two-step approach to predictive modeling using regularization methods including the Lasso family and the ridge regression (Hoerl and Kennard (1970)). In Step one, the Lasso algorithm is applied on all training data to produce a sequence of subsets of selected variables by varying the hyperparameter value. In Step two, for each subset of the selected variables from

Step one, the Lasso, elastic-net (Zou and Hastie (2005)), adaptive Lasso (Zou (2006)), or ridge regression is carried out with cross-validation to find the optimal model among all subsets of selected variables. We demonstrate using industrial datasets that the two-step methods can yield significantly better models than the standard Lasso and other one-step methods.

2. SPARSE LEARNING METHODS AND REGULARIZATION

2.1 The Lasso and ridge regression

Let $\boldsymbol{x}_k = [x_1,\ x_2, \cdots, x_p]^\top \in \Re^p$ be the process variables to be selected and y_k be the response variable to be predicted. For convenience these variables are scaled to zero mean and unit variance based on N samples in the training set. We try to estimate the regression coefficients based on

$$y_k = \beta_0 + \boldsymbol{x}_k^\top \boldsymbol{\beta} + \varepsilon_k, \qquad (1)$$

where ε_k is the zero-mean random noise and $\beta_0 = 0$ if the training data $\boldsymbol{x}_k$ and y_k are scaled to zero mean. The Lasso approach adopts the least squares objective with an l_1 norm penalty of the coefficients as

$$\hat{\boldsymbol{\beta}}_\lambda^{La} = \arg\min_{\boldsymbol{\beta},\beta_0} \frac{1}{2N} \sum_{k=1}^{N} (y_k - \beta_0 - \boldsymbol{x}_k^\top \boldsymbol{\beta})^2 + \lambda \|\boldsymbol{\beta}\|_1 \qquad (2)$$

where λ is the hyperparameter to be tuned optimally based on cross validation.

The typical cross validation process consists of three steps: i) divide the training data into multiple folds with similar distributions and build Lasso models on all but one fold of the data for a grid of λ values; ii) use the trained model to calculate the mean squares error predicted (MSEP) on the fold of data not used for training. Average the MSEPs of all folds for each λ value to generate a sequence of `MSEP` vs. λ; and iii) select the optimal λ value that yields the minimum `MSEP` and use it to re-run Lasso on all training data to obtain the final model, which gives the estimates of non-zero coefficients and zero coefficients that enable variable selection.

The Lasso objective (2) is a variant of the well-known ridge regression (RR) objective which uses an l_2 norm penalty as

$$\hat{\boldsymbol{\beta}}_\lambda^{RR} = \arg\min_{\boldsymbol{\beta},\beta_0} \frac{1}{2N} \sum_{k=1}^{N} (y_k - \beta_0 - \boldsymbol{x}_k^\top \boldsymbol{\beta})^2 + \frac{1}{2}\lambda \|\boldsymbol{\beta}\|_2^2 \qquad (3)$$

As explained in James et al. (2013), the difference between the Lasso and ridge regression solutions can be illustrated in Figure 1, where the elliptic contours represent the least squares error surface, while the diamond and circle represent the effects of the Lasso l_1 and ridge l_2 constraints, respectively. The elongated ellipses illustrate the presence of collinearity among the process variables. The optimal solution is where the constraint regions are tangent to one of the ellipses. It is clear that the Lasso solution makes one coefficient $\hat{\beta}_2 = 0$, which effectively eliminates Variable 2 from the model. On the other hand, the ridge regression solution gives both coefficients $\hat{\beta}_1 \neq 0$ and $\hat{\beta}_2 \neq 0$, which does not eliminate any variables.

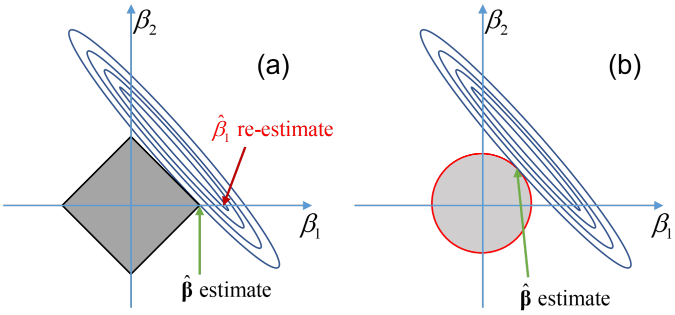
β_2
(a)
$\hat{\beta}_1$ re-estimate
β_1
$\hat{\boldsymbol{\beta}}$ estimate
β_2
(b)
β_1
$\hat{\boldsymbol{\beta}}$ estimate

Fig. 1. (a) Lasso and (b) ridge regression least squares error surface with a diamond shape constraint from an l_1 norm and a circle shape constraint from an l_2 norm. The elongated ellipses represent the effect of collinearity among the process variables.

Some drawbacks of the Lasso solution can be explained with the figure, especially in the case of collinear variables. In Figure 1(a), a minor change in the training data can make the ellipses less tilted so that the optimal solution switches to the vertex at the top of the diamond region, which leads to $\hat{\beta}_1 = 0$ and thus Variable 1 to be eliminated. This is the instability issue of the Lasso discussed in Qin and Liu (2021). On the other hand, the ridge regression solution shown in Figure 1(b) is little changed and thus robust to such a perturbation.

The sub-optimality of the Lasso which uses one hyper-parameter value to yield zero coefficients and estimates of the non-zero coefficients can also be illustrated with Figure 1(a). Under the condition that $\hat{\beta}_2 = 0$ and thus Variable 2 is eliminated from the model, a better estimate of $\hat{\beta}_1$ can exist that achieves a lower error value, which is indicated as the $\hat{\beta}_1$ re-estimate in Figure 1(a). The $\hat{\beta}_1$ re-estimate finds the lowest error value along the β_1 direction, which is a relaxed solution.

2.2 Relaxed Lasso

A relaxed Lasso (Relaxo) approach is proposed by Meinshausen (2007) to overcome the above problem. The Relaxo first uses the Lasso to perform variable selection by specifying $\lambda = \lambda_i$ in (2), then it relaxes the l_1 penalty by introducing a parameter $0 < \phi \leq 1$ and solves a Lasso problem with the selected variables in the first step as

$$\hat{\boldsymbol{\beta}}_{i,\phi}^{RL} = \arg\min_{\boldsymbol{\beta},\beta_0} \frac{1}{2N} \sum_{k=1}^{N} (y_k - \beta_0 - \boldsymbol{\beta}^\top \boldsymbol{x}_k^{(i)})^2 + \lambda_i \phi \|\boldsymbol{\beta}\|_1 \qquad (4)$$

where $\boldsymbol{x}_k^{(i)}$ is the subset of variables selected by the standard Lasso. When $\phi = 1$, both Lasso and Relaxo are the same. For $\phi < 1$, a solution with less penalty is implemented on the selected variables from the first step. This is a kind of de-biasing which makes the solution closer to the unbiased least squares solution.

2.3 Elastic-net

Zou and Hastie (2005) proposed the elastic-net which uses a combination of the Lasso and ridge penalties to improve the stability in selecting variables. The method applies both l_1 and l_2 norms to minimize the following objective function

$$\min_{\boldsymbol{\beta},\beta_0} \frac{1}{2N}\sum_{k=1}^{N}(y_k - \beta_0 - \boldsymbol{x}_k^\top \boldsymbol{\beta})^2 + \lambda\left(\frac{1-\alpha}{2}\|\boldsymbol{\beta}\|_2^2 + \alpha\|\boldsymbol{\beta}\|_1\right) \tag{5}$$

For a given λ, the elastic-net penalty hybridizes the diamond and circular shaped constraints in Figure 1 to form quadratic arcs, which retain vertices to select variables and are strictly convex. The second parameter α tunes the convexity of the arcs to make it difficult to switch between vertices, thus, alleviating the stability issue.

2.4 Adaptive Lasso

One desirable property for variable selection is its consistency. The consistency in variable selection is concerned with whether the estimated zero coefficients are indeed zero in the true model that generates the data. Zou (2006) shows that the Lasso can be inconsistent in variable selection under certain conditions. To make the Lasso consistent, an adaptive Lasso (AdaLasso) algorithm is proposed, which uses the following objective

$$\min_{\boldsymbol{\beta},\beta_0} \frac{1}{2N}\sum_{k=1}^{N}(y_k - \beta_0 - \boldsymbol{x}_k^\top \boldsymbol{\beta})^2 + \lambda\sum_{j=1}^{p}\hat{w}_j|\beta_j|_1 \tag{6}$$

where $\boldsymbol{\beta} = [\beta_1 \quad \cdots \quad \beta_p]^\top$ the weight $\hat{w}_j$ is adaptive for each variable. AdaLasso suggests that $\hat{w}_j = |\hat{\beta}_j^{-1}(\texttt{ols})|$ is a good choice, where $\hat{\beta}_j(\texttt{ols})$ is the solution for the j^{th} variable from the ordinary least squares. The adaptive weights effectively apply more penalty on coefficients that have smaller magnitudes from the least squares estimates.

3. TWO-STEP SPARSE LEARNING FOR VARIABLE SELECTION AND ESTIMATION

In Lasso, the λ value that drives the appropriate variable selection is the same λ value that regularizes the estimates of the nonzero model coefficients. However, the two tasks do not necessarily share the same optimal λ value. First, there is the λ for the l_1 penalty that leads to the optimal selection of variables. Once the subset of variables is selected, one should re-tune the hyperparameter to achieve optimally regularized estimates for the non-zero coefficients. This second step can be achieved with other regularization methods, such as the Lasso, ridge regression, elastic-net, or AdaLasso, which leads to the following two-step approach. The two-step sparse learning can be viewed as a generalization of the Relaxo, which goes beyond the l_1 penalty.

Algorithm 1. Two-Step Sparse Learning with Cross Validation

(1) Scale all training data $\{\boldsymbol{x}_k, y_k\}_{k=1}^N$ to zero mean and unit variance.
(2) **Step One**. Use all training data to estimate $\hat{\boldsymbol{\beta}}_\lambda^N$ in (2) for a grid of λ values to generate a regularization path (Hastie et al. (2015)). The selected subsets of variables along the path are denoted as $\boldsymbol{x}_k^{(1)}, \boldsymbol{x}_k^{(2)}, \cdots, \boldsymbol{x}_k^{(m)}$ without repetition, where $\boldsymbol{x}_k^{(i)}$ contains the selected variables with nonzero coefficients.
(3) **Step Two**. Divide the training data into s fold to perform a regularized regression algorithm with cross-validation, such as the Lasso, ridge regression, elastic-net, or AdaLasso. Estimate the j^{th} CV model using the training set $\mathcal{T}_j$ with N_j observations and the rest as the j^{th} validation set $\mathcal{V}_j$, where $\mathcal{T}_j$ includes all observations except for $\mathcal{V}_j$. For example, if the Lasso is adopted in the second step, the modified Lasso objective for $\boldsymbol{x}_k^{(i)}, i = 1, 2, \cdots, m$ is

$$\hat{\boldsymbol{\beta}}_{i,j,\mu} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2N_j}\sum_{k\in\mathcal{T}_j}(y_k - \beta_0 - \boldsymbol{\beta}^\top \boldsymbol{x}_k^{(i)})^2 + \mu\|\boldsymbol{\beta}\|_1 \tag{7}$$

For each i tune μ for a grid of values to find the $\mu = \mu_i^*$ that yields the best mean squared error predicted (MSEP) by s-fold cross-validation. Denote the optimal MSEP as $\texttt{MSEP}_{i,\mu_i^*}$.
(4) Let $i^* = \arg\min_i \texttt{MSEP}_{i,\mu_i^*}$ among $i = 1, 2, \cdots, m$. The i^*-th subset of variables $\boldsymbol{x}_k^{(i^*)}$ is the optimally selected variables. The corresponding final model coefficients $\hat{\boldsymbol{\beta}}_{i^*,\mu_i^*}$ is re-estimated using all training data with the sparse learning method used in Step Two.

Step One of the Algorithm uses the l_1 penalty to find a sequence of subsets of selected variables along the regularization path, then Step Two determines which subset of variables among all is the optimal one by another regularization method with cross-validation. Since variable selection is not an ultimate concern in Step Two, ridge regression among other regularization methods can be used to estimate the non-zero coefficients. The two-step methods are referred to as *Lasso-Lasso*, *Lasso-ridge* , *Lasso-AdaLasso*, or *Lasso-AdaRidge*, where Lasso-AdaRidge uses the AdaLasso weights in the second Ridge step to implement adaptive ridge regression. Note that the regularization adopted in the second step can handle potential collinearity among the selected subsets of variables.

4. EXPERIMENTAL RESULTS ON A NOX EMISSION DATASET

The industrial boiler data has nine process variables and one output variable, which is the NOx concentration near the top of the stack. The dataset used for this study has 390 observations sampled at a 5-minute interval, of which the detail can be found in Qin and Liu (2021). This boiler operation data are highly collinear, which can be interpreted with the physical process knowledge.

We perform seven-fold cross validation to select optimal hyperparameters. We choose to divide the data into 14 consecutive blocks and then split the blocks into seven folds. The model quality is measured by the overall MSEP and Q^2, which are calculated as follows.

(1) Calculate the sum of squared errors predicted, $\texttt{SSEP}_j$, for the j-th fold.
(2) Sum up $\texttt{SSEP}_j$ to obtain the SSEP and divide it by the total number of samples to obtain the overall MSEP.
(3) This overall SSEP is used to calculate the overall cross-validated R^2, which is denoted as Q^2 and given as follows,

$$Q^2 = 1 - \frac{\mathtt{SSEP}}{\sum_{k=1}^{N}(y_k - \bar{y})^2}$$

where $\bar{y}$ is the mean of the training data, which is zero after the data is scaled to zero mean.

4.1 NOx emission predictions with Lasso

We first establish the baseline modeling result using the standard Lasso Tibshirani (1996). Figure 2 shows two results in one chart. One is the MSEP of the seven-fold cross validation with the Lasso algorithm, shown in circles. The minimum MSEP achieved by the Lasso model establishes a baseline for comparison with the proposed methods. The other one, shown in solids dots, is the number of nonzero coefficients as λ varies by fitting the Lasso model to all data. A set of unique model structures is results along the regularization path. For each of the candidate model structures we search for the optimal $\mu^*(\lambda_i)$ with cross-validation in a subsequent Lasso or ridge regression step.

The Lasso model with the cross-validated MSEP in Figure 2 yields two local minimum solutions, which are denoted as λ_1^* and λ_2^*. λ_1^* picks eight variables, while λ_2^* picks four variables. It is well known that the Lasso tends to pick excessive number of variables.

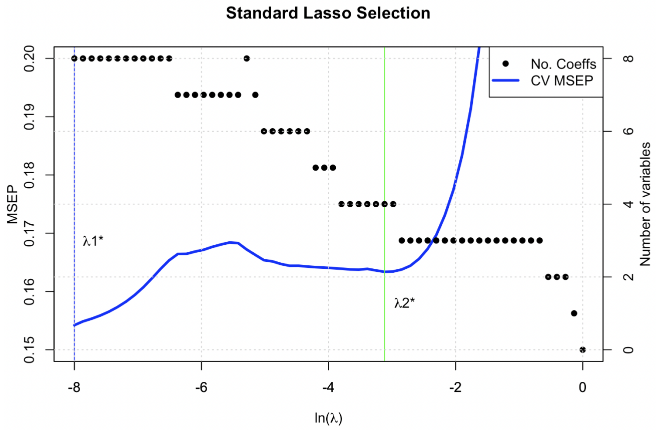

Fig. 2. Cross validated MSEP, shown in circles, with the Lasso algorithm. Two local minima, λ_1^* and λ_2^*, are picked. Solid dots show the number of non-zero coefficients for the selected variables along the regularization path.

To help visualize the unique model structures along the Lasso regularization path, we show in Figure 3 the selected variables as λ varies from small to large, by fitting the Lasso model to all data. It is observed that in some regions the model structure is stable for a wide region of λ, while in a few regions the structures change quickly. The smallest λ_i that leads to a new model structure will be useful for subsequent Lasso-Lasso and Lasso-ridge modeling.

4.2 NOx emission predictions with Lasso-Lasso

To perform Lasso-Lasso modeling, we begin with a subset of selected variables obtained by varying λ_i along the Lasso regularization path. Then we perform cross validation with the second Lasso step by varying the regularization parameter $\mu(\lambda_i)$ between 0 and λ_i. We end up with an MSEP curve for each λ_i, which is depicted in the top panel of Figure 4. We find the $\mu^*(\lambda_i)$ that leads to the minimum MSEP among all curves, which is shown in the bottom panel of Figure 4. In both panels the MSEP of Lasso is shown as blue lines for comparison. From the bottom panel of Figure 4 we see that i) the cross-validated MSEPs obtained by the Lasso-Lasso is much smaller than that obtained by the Lasso; ii) the minimum MSEP reached by the Lasso-Lasso uses λ_i that selects one predictor only, which is Windbox Pressure. The optimal $\mu^*(\lambda_i)$, being much smaller than the λ_i, indicates that the Lasso-Lasso method unleashes the potential of achieving a much smaller MSEP.

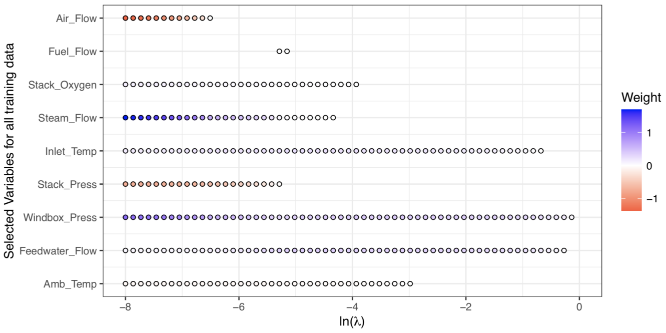

Fig. 3. The selected variables as λ varies from small to large, by fitting the Lasso model to all data. The color bar indicates the signs and magnitude of the coefficients.

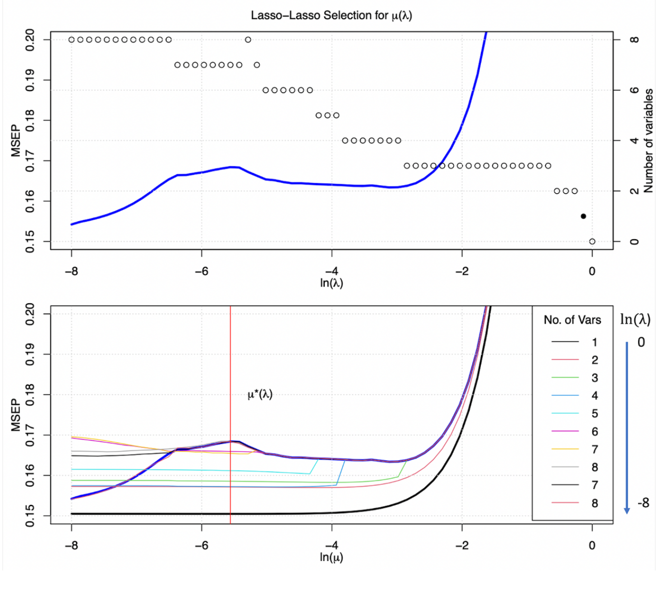

Fig. 4. The cross-validated MSEP's of the Lasso-Lasso: (Top) MSEP's vs. $\mu(\lambda_i)$ for the model structures determined by λ_i in Step 1 of Lasso-Lasso; (Bottom) MSEP's for the λ_i that yield the optimal $\mu^*(\lambda_i)$. The legends show the number of variables in each unique subset of selected variables.

We show in Figure 5 the predictions of the *validation* subsets of the data during cross validation. The top two panels are for the Lasso with optimal λ_1^* and λ_2^*, respectively. The third panel is the prediction of using one variable, which leads to much better predictions with the optimized $\mu^*(\lambda_i)$ from the second Lasso.

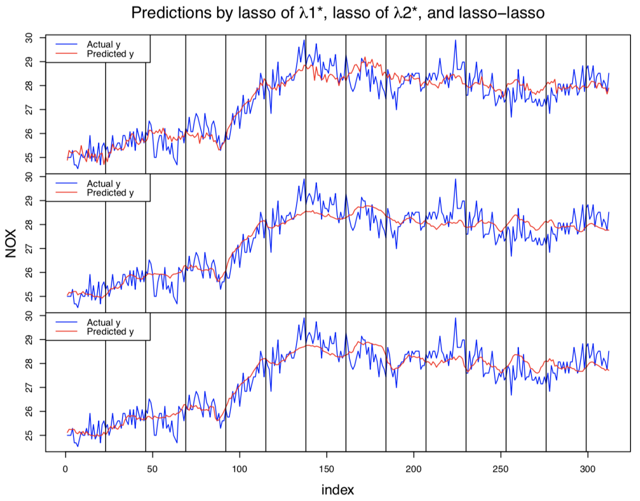

Fig. 5. Predictions of the *validation* folds of the data during cross validation. The top two panels are from the two Lasso solutions, while the bottom panel is the result using one selected variable in the Lasso-Lasso solution.

4.3 NOx emission predictions with Lasso-ridge

The Lasso-ridge modeling procedure uses ridge regression in the second regularization step to find the optimal parameter $\mu^*(\lambda_i)$. The ridge regression uses the selected variables in a candidate subset determined along the Lasso regularization path. The multiple `MSEP` curves due to multiple λ_i are depicted in the top panel of Figure 6. The $\mu^*(\lambda_i)$ that yields the minimum `MSEP` is shown in the bottom panel of Figure 6. In both panels the `MSEP` of the Lasso is shown for comparison. From the bottom panel of Figure 6 we see that the MSEPs obtained by Lasso-ridge is much smaller than those obtained by Lasso.

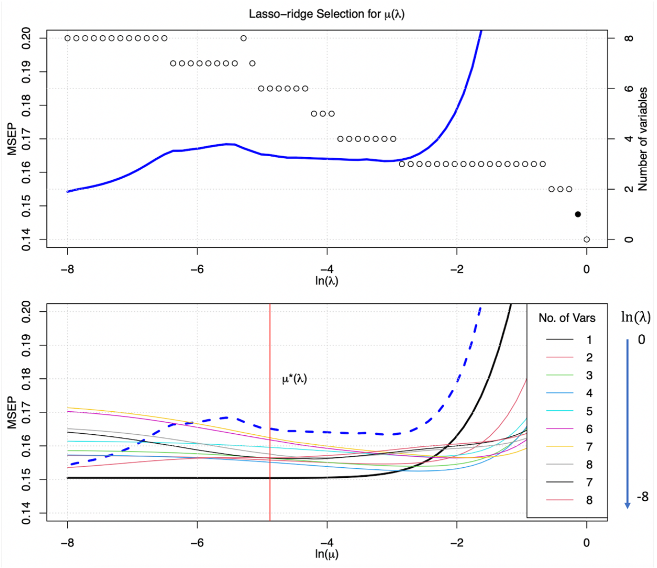

Fig. 6. The `MSEP`'s of Lasso-ridge: (Top) `MSEP` curves of Lasso-ridge vs. $\mu(\lambda_i)$; (Bottom) Lasso-ridge `MSEP`'s by the λ_i that leads to the optimal $\mu^*(\lambda_i)$. The legends show the number of variables in each unique subset of selected variables.

Table 1. MSEPs and Q^2 indices of the optimal Lasso models, the Lasso-Lasso, and the Lasso-ridge

	Lasso, λ_1^*	Lasso, λ_2^*	Lasso-Lasso	Lasso-ridge
`MSEP`	0.1523	0.1629	**0.1485**	**0.1485**
Q^2	0.847	0.837	**0.851**	**0.851**

Figure 7 shows the Lasso-ridge predictions of the *validation* subsets of the data during cross validation. The top panel shows the predictions with the optimal Lasso model, while the bottom panel shows the optimized Lasso-ridge result of using that one predictor. Lasso prevents itself from reaching a much better model that is achieved by Lasso-ridge method. The cross-validated MSEPs will be summarized in the next subsection for comparison.

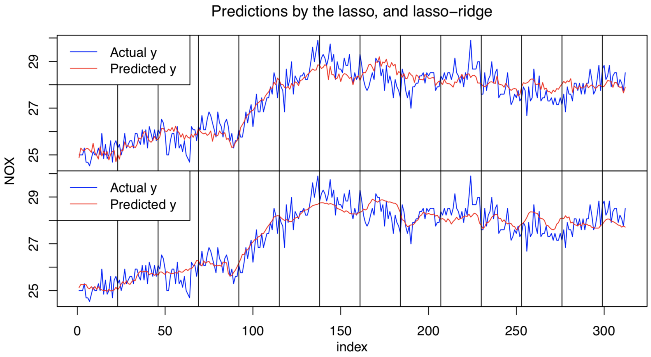

Fig. 7. Predictions of the *validation* folds of the data during cross validation. The top and bottom panels show the predictions of the Lasso and Step Two of the Lasso-ridge, respectively.

4.4 Comparing Lasso-Lasso and Lasso-ridge on the NOx emission data

The `MSEP`'s and Q^2 indices of the optimal Lasso models with λ_1^* and λ_2^*, the Lasso-Lasso model and the Lasso-ridge model are shown in Table 1. The results show the Lasso-Lasso and Lasso-ridge models give the best cross-validation errors, which selects Windbox Pressure as the only predictor. The other variables are highly correlated with Windbox Pressure for the boiler process.

To examine the difference between the Lasso-Lasso and the Lasso-ridge models, we show in Figure 8 the minimum MSEP's achieved by the Lasso-Lasso models and the Lasso-ridge models from the candidate model structures selected by the Lasso. It is observed that the Lasso-ridge models achieve lower `MSEP` than the Lasso-Lasso models for the same candidate model structure.

The regression coefficients from the Lasso with λ_1^*, the Lasso with λ_2^*, the Lasso-Lasso, and the Lasso-ridge models are given in Table 2. It is observed from the table that the Lasso model with λ_1^* yields negative coefficients for two variables, Air Flow and Stack Pressure. Since most of the variables are positively correlated, it contradicts the first principles that these coefficients are negative. Therefore, the Lasso model with λ_1^* is not physically interpretable.

In some works it is suggested to determine the optimal subset of variables via the Lasso as Step One, and then

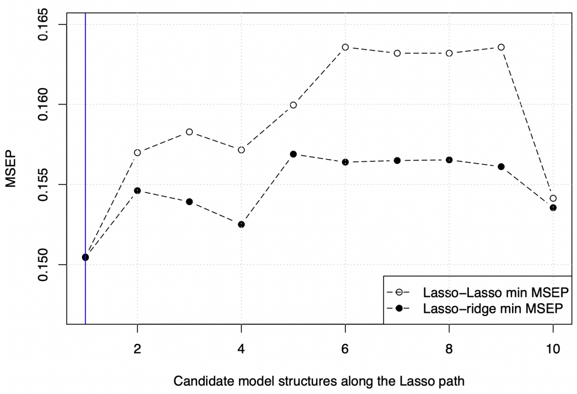

Fig. 8. Minimum MSEP's achieved by Lasso-Lasso and Lasso-ridge based on the subsets of selected variables by the Lasso.

Table 2. Regression coefficients of the two optimal Lasso models, the Lasso-Lasso, and the Lasso-ridge models

	Lasso, λ_1^*	Lasso, λ_2^*	Lasso-Lasso	Lasso-ridge
(Intercept)	0	0	0	0
Air Flow	-1.5203	0	0	0
Fuel Flow	0	0	0	0
Stack Oxygen	-0.1160	0	0	0
Steam Flow	1.7624	0	0	0
Inlet Temp	0.1390	0.1889	0	0
Stack Press	-0.7095	0	0	0
Windbox Press	1.1855	0.3657	0.9216	0.9184
Feedwater Flow	0.0687	0.3345	0	0
Ambient Temp	-0.0577	-0.0086	0	0

use ordinary least squares to estimate the non-zero parameters. One question is in order: could such an approach find the optimal solutions that are found by Lasso-ridge and Lasso-Lasso? The answer is no. As illustrated in Table 2, the optimal model structure by the Lasso has either 8 or 4 variables, whereas both Lasso-ridge and Lasso-Lasso find one predictor to be optimal.

4.5 Monte-Carlo simulation of various sparse learning methods

To test how the various sparse learning methods perform on the NOx emission data, we perform 20 Monte-Carlo simulations by randomly selecting the 14 segments of data into seven-fold. The randomness allows us to see the variations among the modeling results. We test all eight methods, Lasso, elastic-net, AdaLasso, Relaxo, Lasso-Lasso, Lasso-ridge, Lasso-AdaLasso, and Lasso-AdaRidge and depict the MSEPs and Q^2 results in Figure 9. For the simple NOx data the adaptive Lasso and Lasso-AdaLasso outperform others. The Lasso and elastic-net incur the largest MSEPs, while the Lasso-ridge reduces the MSEP of the Lasso with the second ridge step. We next apply the methods to the Dow challenge dataset to compare the effectiveness of the methods.

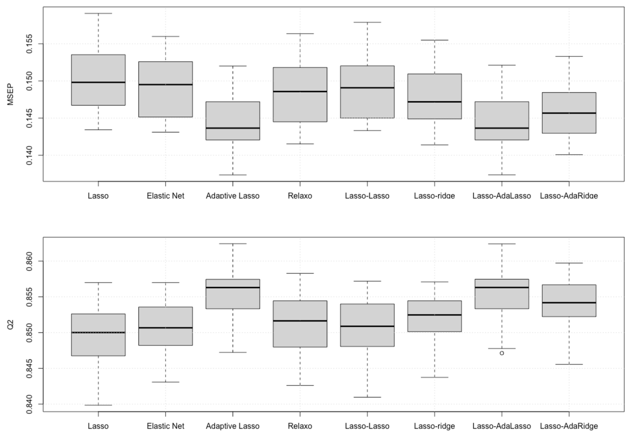

Fig. 9. Boxplots of 20 Monte-Carlo MSEP and Q^2 for each method.

5. TEST ON THE DOW DATA CHALLENGE PROBLEM

The Dow data challenge problem was posted by Dow Braun et al. (2020) to help academia test their algorithms with an industrially relevant problem. Measurements of 44 process variables are available for variable selection to build the best model to predict impurity in the product stream. In this paper, we use the training dataset that was pre-processed by Joe Qin et al. (2021), where interpolated impurity data were removed and the impurity was transformed by a softplus function to guarantee positive impurity predictions.

Due to limited space allowed for this paper, we only show the cross-validation results in Step Two. Figure 10 shows the predictions of the *validation* folds of the data during cross validation. The panels show the predictions vs. actual data for each fold when it is used as a validation fold with the Lasso, elastic-net, AdaLasso, Relaxo, Lasso-Lasso, Lasso-ridge, Lasso-AdaLasso, and Lasso-AdaRidge, respectively, from top to bottom. It is easy to observe in the first and last two folds, some methods fail to predict the impurity accurately.

The Lasso-ridge and Lasso-AdaRidge seem to give the best cross-validated predictions compared to the actual data.

The cross-validated MSEPs and Q^2 indices of the optimal models by the eight methods are shown in Figure 11, which clearly shows that the Lasso-ridge and Lasso-AdaRidge give the best models with smallest cross-validation errors. While adaptive Lasso does well for the NOx data, ridge regression based methods are the winner for the Dow impurity dataset.

More observations from Figure 10 are in order. The elastic-net includes all variables, which finds $\alpha = 0.001$ to be optimal. This is essentially a ridge regression model. The AdaLasso and Lasso-AdaLasso estimate of the coefficients are very large magnitudes, which can inflate the prediction variance. The Lasso-Lasso uses the fewest number of variables. Finally, the Lasso-AdaRidge is not better than Lasso-ridge, even with its adaptive weight scheme.

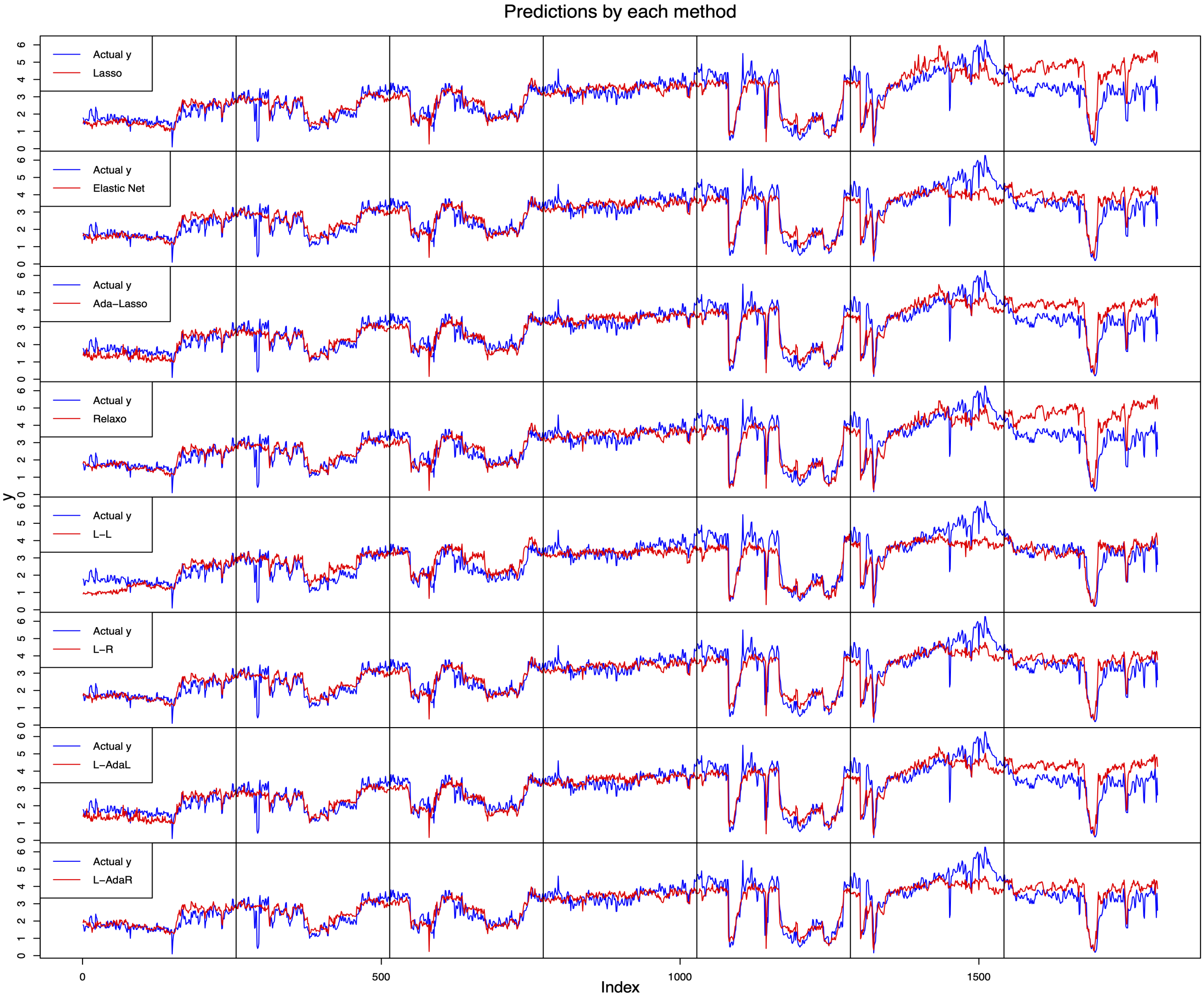

Fig. 10. Predictions of the *validation* folds of data during cross validation from the Lasso, elastic-net, AdaLasso, Relaxo, Lasso-Lasso, Lasso-ridge, Lasso-AdaLasso, and Lasso-AdaRidge, respectively.

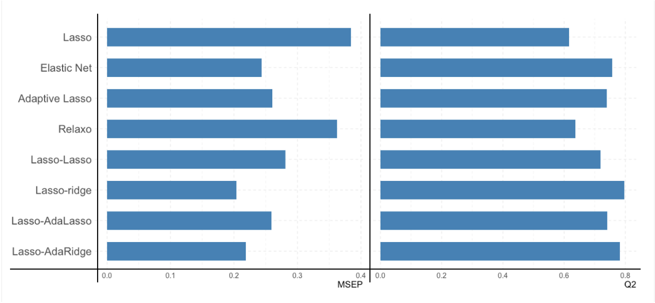

Fig. 11. Dow impurity MSEPs and Q2 indices achieved by the optimal models of the various sparse learning methods.

6. CONCLUSIONS

A novel two-step sparse learning approach is proposed in this paper to select variables and estimate model parameters optimally. The proposed algorithms, including *Lasso-Lasso*, *Lasso-ridge*, *Lasso-AdaLasso*, and *Lasso-AdaRidge*, can be viewed as extensions of the relaxed Lasso proposed by Meinshausen (2007). The paper reveals that the Lasso tends to over constrain itself from reaching a better estimate of the non-zero coefficients due to the use of one hyperparameter for the two problems. This observation is demonstrated with industrial datasets of the Dow data challenge problem and a boiler NOx prediction problem. The elastic-net does outperform other methods for the NOx emission data, while for the Dow data it does not remove variables, leading to essential a ridge regression solution.

It is further observed with the two industrial application studies that the adaptive Lasso, Lasso-ridge and Lasso-AdaRidge give better predictions than other methods that use l_1 norm in the second step. One evidence of using the Lasso l_1 norm penalty is that it tends to remove all but one variable that has a marginally higher correlation with the output than other variables.

ACKNOWLEDGMENTS

Financial support for this work from the City University of Hong Kong under Project 9380123 and an NSF-China Regional Joint Key Project for Innovations and Development (U20A20189) is gratefully acknowledged.

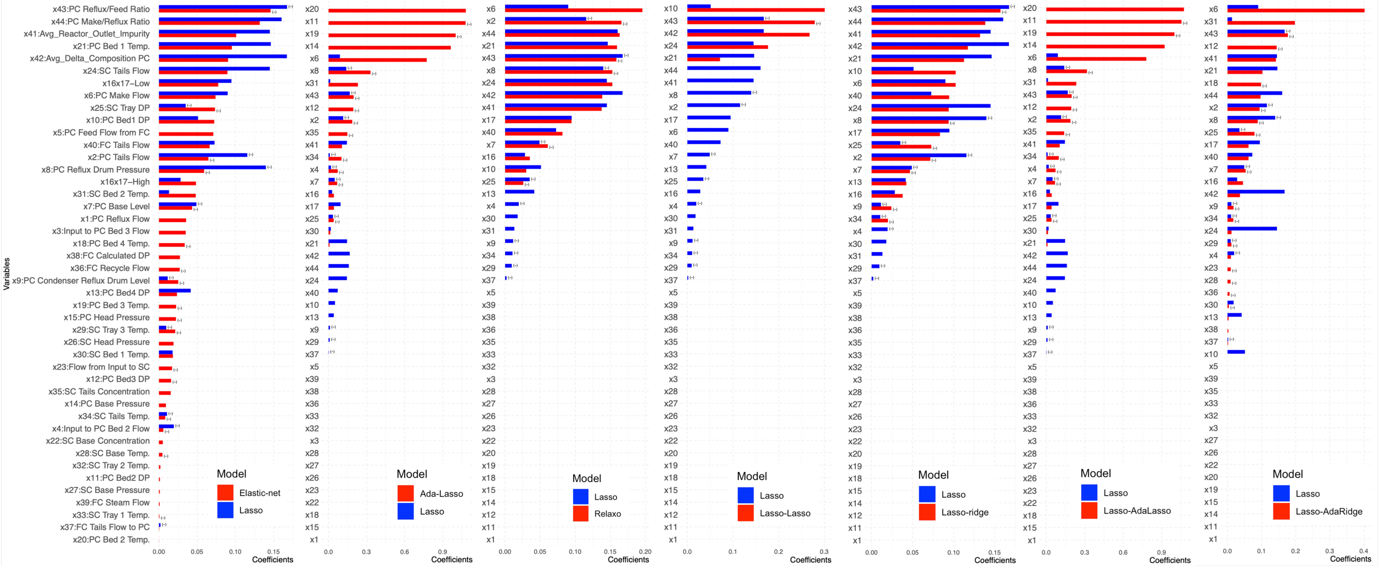

Fig. 12. Estimated coefficients for the Dow impurity data with various sparse learning methods.

REFERENCES

Arora, N. and Kaur, P.D. (2020). A Bolasso based consistent feature selection enabled random forest classification algorithm: An application to credit risk assessment. *Applied Soft Computing*, 86, 105936.

Braun, B., Castillo, I., Joswiak, M., Y. Peng, R.R., Schmidt, A., Wang, Z., Chiang, L., and Colegrove, B. (2020). Data science challenges in chemical manufacturing. In *IFAC World Congress Proceedings*. Berlin, Germany.

Galicia, H.J., He, Q.P., and Wang, J. (2011). A reduced order soft sensor approach and its application to a continuous digester. *Journal of Process Control*, 21(4), 489–500.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical Learning with Sparsity: the Lasso and Generalizations*. CRC press.

Hoerl, A. and Kennard, R. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12, 55–67.

James, G., Witten, D., Hastie, T., and Tibshirani., R. (2013). *An Introduction to Statistical Learning: with Applications in R*. Springer.

Joe Qin, S., Guo, S., Li, Z., Chiang, L.H., Castillo, I., Braun, B., and Wang, Z. (2021). Integration of process knowledge and statistical learning for the Dow data challenge problem. *Computers & Chemical Engineering*, 107451.

Kamkar, I., Gupta, S.K., Phung, D., and Venkatesh, S. (2015). Stable feature selection for clinical prediction: Exploiting ICD tree structure using tree-lasso. *Journal of Biomedical Informatics*, 53, 277 – 290.

Kano, M., Miyazaki, K., Hasebe, S., and Hashimoto, I. (1998). Inferential control system of distillation compositions using dynamic partial least squares regression. *IFAC Proceedings Volumes*, 31, 375–384.

Khatibisepehr, S., Huang, B., and Khare, S. (2013). Design of inferential sensors in the process industry: A review of Bayesian methods. *Journal of Process Control*, 23(10), 1575–1596.

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

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4), 417–473.

Qin, S.J., Yue, H., and Dunia, R. (1997). Self-validating inferential sensors with application to air emission monitoring. *Industrial & Engineering Chemistry Research*, 36, 1675–1685.

Qin, S.J. and Liu, Y. (2021). A stable Lasso algorithm for inferential sensor structure learning and parameter estimation. *Journal of Process Control*, 107, 70–82.

Qin, S. and McAvoy, T. (1992). A data-based process modeling approach and its applications. *IFAC Proceedings Volumes*, 25(5), 93–98.

Shang, C., Yang, F., Huang, D., and Lyu, W. (2014). Data-driven soft sensor development based on deep learning technique. *Journal of Process Control*, 24(3), 223–233.

Sun, W., Wang, J., and Fang, Y. (2013). Consistent selection of tuning parameters via variable selection stability. *The Journal of Machine Learning Research*, 14(1), 3419–3440.

Tham, M.T., Montague, G.A., Morris, A.J., and Lant, P.A. (1991). Soft-sensors for process estimation and inferential control. *Journal of Process Control*, 1(1), 3–14.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58(1), 267–288.

Zhao, H. (2021). An industry perspective on AI, machine learning and data science towards industry 4.0. In *Workshop Series on Control Systems and Data Science towards Industry 4.0*.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, 67(2), 301–320.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476), 1418–1429.