LEAST ANGLE REGRESSION AND PARTIAL LEAST SQUARES REGRESSION ON PROCESS DATA WITH HIGH COLLINEARITY

Siyi Guo[1], Kenmond Pang[1] and S. Joe Qin[1]

[1]Mork Family Department of Chemical Engineering & Materials Science, University of Southern California,
3650 McClintock Ave, Los Angeles, CA 90089

Abstract Overview

Collinearity is a common problem met in regression analysis for chemical process data. A popular method to deal with collinearity is partial least squares regression (PLS). It uses projections of the original variables to a reduce number of latent variables to circumvent the task of model selection. On the other hand, recent advances in statistics and machine learning provide promising methods of sparse analytics, which lead to a natural way to exclude variables that are irrelevant or redundant. To study the effectiveness of these methods, we evaluate the performance of least angle regression (LARS) on an industrial boiler dataset with high collinearity, and compare its performance with those of least absolute shrinkage and selection operator (LASSO) and PLS. The results show that LARS has a better performance on highly collinear data. It produces sparse coefficients like LASSO, which PLS cannot achieve, and it allows for an easier selection of a best set of coefficients comparative to LASSO.

Keywords

Least Angle Regression, LARS, Partial least squares, PLS, Collinearity.

Introduction

Collinearity, or the high correlation among independent predictor variables, is often met in chemical process data analytics. It causes variance inflation and therefore hampers the accuracy of parameter estimation. Many methods have been studied in this aspect. PLS deals with multicollinearity by projecting the dataset onto a latent space and selecting the latent variables which explains the largest amount of variance in the observed variables. Another method for addressing collinearity is least absolute shrinkage and selection operator (LASSO), which performs variable selection by penalizing the coefficient estimation. Most recently, the least angle regression (LARS) seems to be a promising tool for variable selection.

Least angle regression is a model selection algorithm that is less greedy and more computationally efficient compared to classic model-selection tools. The algorithm is as following: starting with all coefficients equal to zero, LARS first finds the predictor most correlated with the response and proceeds in the direction of this predictor until some other predictor has as much correlation with the current residual. Next, LARS goes in a direction equiangular between the two predictors, or along the "least angle direction", until a third predictor has as much correlation with the current residual. Then LARS proceeds equiangularly among these three predictors. The algorithm keeps proceeding this way until all the predictors have non-zero coefficients. LARS can also be easily modified to produce LASSO estimates. The modification is presented in Efron et al. (2004). It is worth studying what difference the LARS-modified LASSO algorithm makes compared to the regular LASSO.

In this paper, we perform LARS and LARS-modified LASSO, along with regular LASSO and PLS on a dataset taken from an industrial boiler process. The dataset contains

eight predictors, which are shown in the diagonal cells in the matrix in Figure 1. The response is the NOx level. Figure 1 shows the high correlations among six predictors.

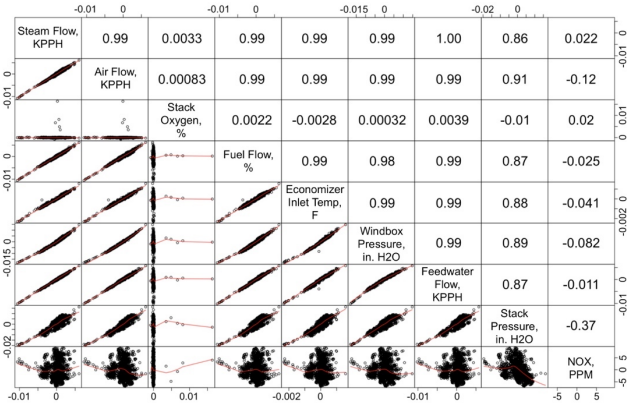

Figure 1. Correlation matrix of the boiler process data

Comparison between LARS and PLS

PLS is a linear regression model by projecting the predictor and response variables onto a new space. It mitigates the impact of collinearity because of its dimensional reduction process. To find out how LARS performs relative to PLS, both methods were implemented on the boiler data using R packages 'pls' and 'lars'. The calculated coefficients are presented in Figure 2. Mean square error of prediction (MSEP) calculated by cross validation is used for error evaluation for PLS results. The minimum MSEP usually corresponds to the best set of coefficients. Cp-type risk estimation is used to assess the error for LARS. Similar to MSEP, the minimum Cp value usually corresponds to the best set of coefficients.

The two methods produce a similar trend of coefficient shrinkage, and they choose the same two variables to have the largest coefficients. However, one obvious advantage of LARS is that it eliminates some predictors by allowing their coefficients to go to zero. PLS on the other hand cannot achieve sparse coefficient estimation. Its coefficient estimates are small for some variables, but they do not go to zero. For LARS the Cp value stops its rapid decrease at step 4, before it reaches the results of ordinary least squares. Even though this Cp value is not the global minimum, it changes very little afterwards. Therefore, it is reasonable to take the coefficients at this step as the best set, in which four variables are selected and all others have zero coefficients. For PLS, the minimum MSEP shows up at 7 components.

Figure 3 shows that steam flow and air flow both have larger coefficients with LARS and PLS. However, for other smaller coefficients, LARS produces four zero coefficients while PLS has all the coefficients to be non-zero. At this point, the PLS result is already very close to ordinary least

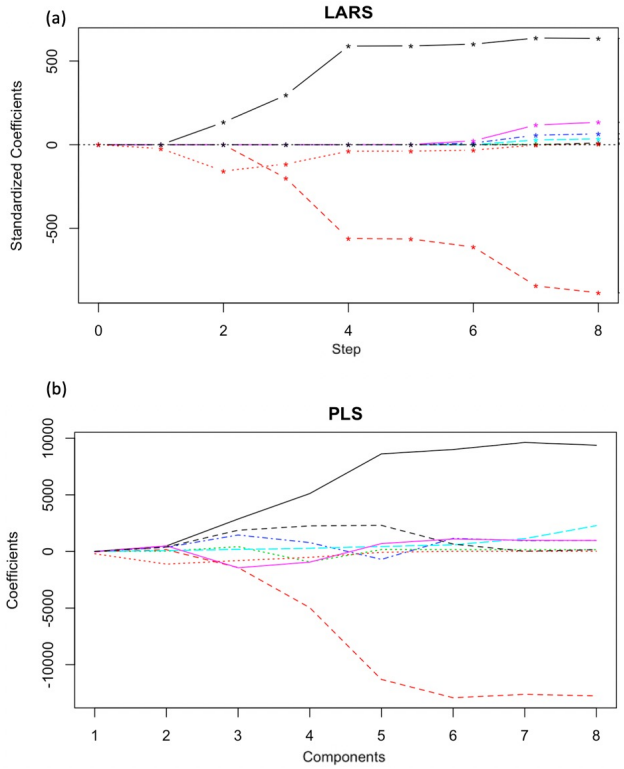

Figure 2. A comparison between LARS and PLS: (a) the standardized coefficient estimates using LARS; (b)the coefficient estimates using PLS

squares. Therefore, even though PLS and LARS both pick steam flow and air flow as leading predictors, PLS does not lend itself to variable selection, since it cannot produce sparse coefficient estimates like LARS does.

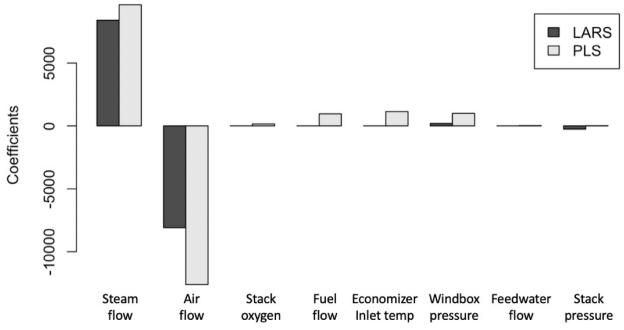

Figure 3. The best sets of coefficients at step 4 for LARS and step 7 for PLS

Comparison between LARS-modified LASSO and The Regular LASSO

The LASSO is a constrained version of ordinary least squares with a penalty term imposed on the L-1 norm of the regression coefficients. Similar to LARS, LASSO is often used as a model selection tool by producing sparse coefficients that have zeros for some variables. LARS-modified LASSO and regular LASSO were performed on

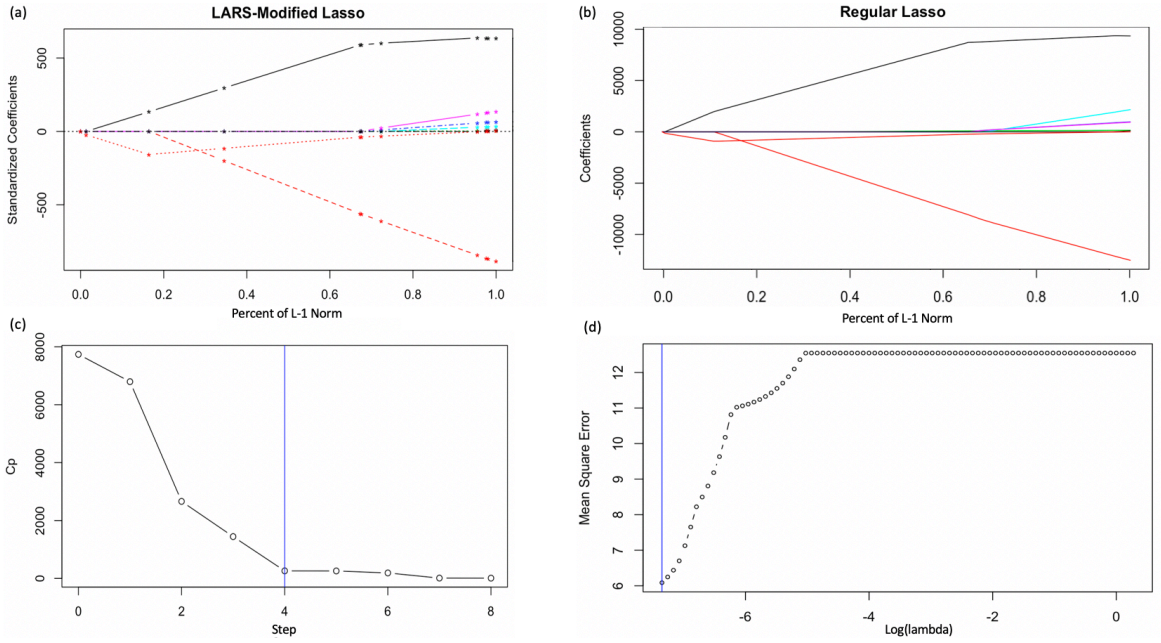

Figure 4. A comparison between LARS-modified and regular LASSO: (a) the standardized coefficient estimates using LARS-modified LASSO; (b)the coefficient estimates using regular LASSO; (c) the Cp value at each step for LARS-modified LASSO; and (d) the MSE at each lambda value in log scale. In (c) and (d), the blue lines indicate the steps of the best set of coefficients.

the boiler data using R packages 'lars' and 'glmnet', respectively. The results are shown in Figure 4.

The coefficient plots are exactly the same between regular and LARS-modified LASSO. However, Figure 4(c) and (d) show a difference in selecting a best set of coefficients. For regular LASSO, this selection is not easily achieved because the minimum MSE appears at a very small λ value, where the coefficients calculated by regular LASSO is very close to those calculated by ordinary least squares. All the predictors have non-zero coefficients in this case. On the other hand, LARS proceeds in the equiangular direction stepwise and takes in one predictor at a time. For LARS-modified LASSO, the Cp value elbow shows up at step 4. Therefore, the coefficients at step 4 can be taken as the best set. This is a good predictor selection, with four variables selected and all others have zero coefficients.

As can be seen, although LARS-modified LASSO produces same coefficient estimates as regular LASSO, due to the nature of LARS algorithm, that it is performed stepwise by taking in one predictor at a time, LARS-modified LASSO is better than LASSO in selecting the best set of coefficients.

Conclusion

The comparative studies of LARS vs. PLS and LARS-modified vs. regular LASSO on the highly collinear boiler data have demonstrated the advantages of LARS. It produces sparse coefficient estimates and achieves variable selection with highly collinear data. The nature of its stepwise algorithm gives the best estimate of coefficients. Last, in Efron et al (2004), the algorithm of LARS requires only the same order of magnitude of computational effort as ordinary least squares applied to the full variable set. This makes LARS computationally efficient when it comes to large datasets.

All the advantages of LARS make it a promising method in data analytics. LARS and its variants are currently being applied to solve LASSO and elastic net problems in various sparse algorithms, such as sparse principal component analysis (SPCA). Further works will be devoted to compare the geometry of LARS with PLS.

References

Efron, Bradley, Hastie, Trevor, Johnstone, Iain and Tibshirani, Robert. "Least Angle Regression." The Annals of Statistics 32.2 (2004): 407-99.

Tibshirani, Robert. "Regression Shrinkage and Selection Via the Lasso." Journal of the Royal Statistical Society: Series B (Methodological) 58.1 (1996): 267-88.

Wold, Svante, Sjöström, Michael, and Eriksson, Lennart. "PLS-regression: A Basic Tool of Chemometrics." Chemometrics and Intelligent Laboratory Systems 58.2 (2001): 109-30.

Zou, Hui, Hastie, Trevor, and Tibshirani, Robert. "Sparse Principal Component Analysis." Journal of Computational and Graphical Statistics 15.2 (2006): 265-86.