Proposal
Sparse Partial Least Squares Regression Study

Shanglin Zhou 2334229

February 20, 2018

Introduction

Along with the the more and more importance of regression-based modeling of high dimensional data problems, regression analysis is now becoming widely used in many fields, such as social science, biological, and also financial. The first method that comes into mind must be ordinary least square (OLS) [3], which is the simplest and also the oldest method tile now. However, when there is multicollinearity among explanatory variables, OLS could not fit well. Then, principle component regression (PCR) [2] came into sight. It could convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables, and then choose the components with higher variance.

Another effective method that was promoted is partial least squares regression (PLS) [5]. By projecting the predicted variables and the observable variables together to a new space, PLS method proposed a new way to solve the multicollinearity.

As overfitting becomes a problem, sparse approaches are introduced to do variable selection. Lasso [4] becomes very popular in the past two decades, by imposed ℓ_1 norm regularization terms, for the reason of simplicity and efficiency.

In the report, we mainly study Sparse PLS method [1]. Sparse PLS not only owns the capability of feature selection in original X space, but also takes advantage of PCR-like dimension reduction. Sparse PLS also maximize the correlation between the predictor and response, thus overcomes the weakness of PCR methods.

Data

The dataset we are going to use is from site "UCI Machine Learning Repository". Data were recorded from March 2004 to February 2005 (one year). It contains 9358 instances of hourly averaged responses from an array of 5 metal oxide chemical sensors embedded in an Air Quality Chemical Multisensor Device. [1] The dataset represents the longest freely available recordings of on field deployed air quality chemical sensor devices responses. Ground Truth hourly averaged concentrations for CO, Non Metanic Hydrocarbons, Benzene, Total Nitrogen Oxides (NOx) and Nitrogen Dioxide (NO2) and were provided by a co-located reference certified analyzer. It is described in Table 1.

[1] `https://archive.ics.uci.edu/ml/datasets/Air+Quality`.

Table 1: Description of Air Quality data.

Variable	Description
Date	DD/MM/YYYY
Time	HH.MM.SS
True hourly averaged concentration CO	mg/m^3
PT08.S1 (tin oxide) hourly averaged sensor response	nominally CO targeted
True hourly averaged overall Metanic HydroCarbons concentration	$microg/m^3$
True hourly averaged Benzene concentration	$microg/m^3$
PT08.S2 (titania) hourly averaged sensor response	nominally NMHC targeted
True hourly averaged NOx concentration	*ppb*
PT08.S3 (tungsten oxide) hourly averaged sensor response	nominally NO_x targeted
True hourly averaged NO2 concentration	$microg/m^3$
PT08.S4 (tungsten oxide) hourly averaged sensor response	nominally NO_2 targeted
PT08.S5 (indium oxide) hourly averaged sensor response	nominally O_3 targeted
Temperature	$^\circ C$
Relative Humidity	%
AH Absolute Humidity	

Method

In the report, we plan to re-perform the experiments in the paper using SPLS method. The software package used as the implementation of SPLS, is the "spls" package for R language. This package is written by the author of [1] thus could be the best candidate.

Moreover, in order to compare SPLS with other methods, Lasso and Elastic Net are also compared as these two methods also enjoy the capability of dimension shrinkage. Lasso and Elastic Net implementation are from the well known "sklearn" package for Python.

References

[1] Hyonho Chun and Sündüz Keleş. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(1):3–25, 2010.

[2] Ian T Jolliffe. A note on the use of principal components in regression. *Applied Statistics*, pages 300–303, 1982.

[3] Adrien Marie Legendre. *Nouvelles méthodes pour la détermination des orbites des comètes*. F. Didot, 1805.

[4] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[5] H Wold. Estimation of principal components and related models by iterative least squares multivariate analysis 1966.