Aix Marseille School of Economics

Master 2: Econometrics, Big Data and statistics

Automatic model selection methods: Midterm evaluation

Author:
Coraline BEST

Professor: Sullivan Hue

February 15, 2024

Table of contents

1 Presentation of the algorithm 3

2 Advantages of Partial Least Squares regressions 3

3 Disadvantages 4

4 Application of the Partial Least Squares regression 4

5 Conclusion 7

1 Presentation of the algorithm

Partial least squares (PLS) were first introduced by the Swedish statistician Herman O. A. Wold. It is a statistical method used for regression and classification. It is particularly useful when dealing with high dimensional data, multicollinearity, and situations where the number of features is greater than the number of observations. PLS is commonly employed in field such as chemometrics, bioinformatics, and machine learning.

PLS has some relations to the principal component regression and can be viewed as an extension. As PCR, PLS can be used as a dimension reduction method, which aims to reduce the number of variables used to predict a given outcome variable. PLS first identifies a new set of features that are linear combinations of the original features and then fits a linear model via least squares using these M new features. Unlike PCR, PLS identifies these new features in a supervised way, meaning it uses as well the outcome variable Y. The idea is to find directions that help explain both the response and the predictors. Therefore, PLS aims at finding the fundamental relations between the X and Y matrices to model the covariance structures in these two spaces. A PLS model will try to find the multidimensional direction in the X space that explains the maximum multidimensional variance direction in the Y space. The components that are selected give the maximal reduction in covariance of the data.

PLS is based on an iterative algorithm. It is necessary to use it on standardized data. To compute the components I successively regress the X matrix on a column of Y to weights. Then the X matrix is regressed on the weights, to locate the variables of X most correlated between themselves. This gives us the scores. Then I regress the Y matrix on to the scores to get the slope coefficients. The Y matrix is then re regressed on the slope coefficients. This final step ensures that the new components are also the most correlated with Y. From these successive regression I get the components that are the linear combination of the original data, in relation with Y. To perform the regression I then select the number of components by cross-validation or an information criteria to finally predict on the new data.

2 Advantages of Partial Least Squares regressions

Partial Least squares might not consistently outperform alternatives like PCR or ridge regression but possess several advantages that make it an effective tool:

1. Handling multicollinearity: PLS excels in managing multicollinearity, particularly when dealing with highly correlated predictors, or when dealing with cases with more predictors than observations. Its focus on the covariance structure allows it to navigate situations where traditional methods may struggle.
2. Dimensionality reduction: PLS serves as a valuable instrument for dimensionality reduction. By transforming original variables into a compact set of components, it streamlines the complexity of the data, making it more manageable.
3. Effective handling of missing data: PLS exhibits proficiency in handling missing data, surpassing the capabilities of other methods. Its adaptability ensures efficient processing even in the presence of incomplete information.
4. Computational efficiency: The model is easy to fit and implement. All the steps in the algorithm are recursive fittings of linear models. This makes PLS accesible to most computers.

5. Versatility in mutlivariate cases: The versatility of PLS extends seamlessly to multivariate scenarios with various responses. This adaptability makes it suitable for a wide range of analytical challenges.

3 Disadvantages

Partial least squares alos have some drawbacks:

1. Interpretability: The components produced by PLS might lack clear interpretability. Compared to PCA, by taking into account the outcome variables, the relationships within the factors can become very complex and not clear for interpretation.
2. Choice of factors: The factor structure is the strength of the PLS models. Nonetheless, if they are not picked accordingly the model can suffer from the pitfalls it tries to solve. Too many factors will result in an overfitted model, not enough will greatly increase the bias. This trade off is usually solved with cross validation, leaving less observations available for fitting the model.
3. Sensitivity to outliers: Similar to many regression techniques, PLS can be sensitive to outliers, impacting its robustness.
4. Non-linear relationships: PLS assumes linear relationships between predictors and responses, potentially limiting its effectiveness in scenarios characterized by intricate non-linear relationships.

4 Application of the Partial Least Squares regression

To illustrate the PLS regression, I used a macroeconomic dataset [1]. This dataset was used by Maehashi and Shintani 2020 in a forecasting comparison of machine learning models, factor based models and standard econometric model for predicting variables such as industrial production, inflation or real estate prices in Japan.

The data is collected from different sources, mainly the Bank of Japan and Ministry of Economy Trade and Industry. It has 588 monthly observations spanning from January 1974 to December 2022, and it observes 219 variables. The variables can be grouped in 9 categories : Real Output, looking at industrial, retail and services activity measures. Inventories, measuring the stocks of the different sectors of the economy. Investments, including productive, real estate and land and investments. Employment, looking at worked hours, wages and number of employed people. Consumption, collected from purchases and sales data. Firms, tracking the financial condition of companies. Money, Stock Price and Interest Rate. Price Indexes and finally Trade.

In this case I want to fit a model around industrial production in the manufacturing sector, labeled as variable **x2**. To do so I proceed in two parts. First I fit a model on the full data set to see if the PLS method can provide a good in-sample fit. I use this result to understand the composition of the factors. Then I use PLS in a forecasting exercise to predict a quarter ahead (3 months) the industrial production value's. In both cases I will use a univariate and multivariate PLS model, using **x107** or New Job Offers as second outcome variable.

I first study the factors created by the PLS model. Because I have 218 regressors I only look at the first 3 factors, and only the 5 variables having the largest weight inside. Otherwise this process would

[1] Link to the dataset and the list of variables

be too long for not much relevant information. In the univariate model the 5 factors are composed the following way, by order of importance.

	Y = x2	**Y = x2, x107**
Factor 1	Index of Capacity Utilization Ratio (Transport Equipment), Index of Industrial Production (Transport Equipment), Index of Industrial Production (Durable Consumer Goods), Index of Producer's Shipments (Durable Consumer Goods), Index of Capacity Utilization Ratio (Machinery)	Index of Industrial Production (Mining and Manufacturing), Index of Industrial Production (Producer Goods), Index of Producer's Shipments (Producer Goods),Index of Industrial Production (Final Demand Goods), Index of Industrial Production (Producer Goods for Mining and Manufacturing)
Factor 2	Total Number of New Housing Construction Started (Government Housing Loan Corporation), Order Received for Construction (Public), Order Received for Construction (Private), Bank Clearings (Number), Order Received for Construction (Grand Total)	Effective Job Offer Rate (Part time),Index of Regular Workers Employment (All Industries - 30 or more persons),Index of Regular Workers Employment (Real Estate), Index of Regular Workers Employment (Construction), Index of Regular Workers Employment (Wholesale and Retail Trade)
Factor 3	Total Number of New Housing Construction Started (Government Housing Loan Corporation), Sales Volume (Daily Average Tokyo Stock Market First Section), Sales Value (Daily Average Tokyo Stock Market First Section), Index of Sales (Retail),Number of New Passenger Car Registrations and Reports (Total)	Index of Regular Workers Employment (Electricity Gas and Heat Supply), Effective Job Offer Rate (Part time), Index of Producer's Inventory of Finished Goods (Mining and Manufacturing), Index of Regular Workers Employment (Manufacturing), Index of Producer's Inventory of Finished Goods (Investment Goods)

Table 1: Description of the 5 largest weights for the first 3 factors in each model

I see that in the univarite case, factor 1 is a proxy measure for the movements in industrial production. The second factor looks at housing market and the last one could be interpreted as a measure of global activity, mixing housing construction, sales and the volume of exchanged stocks. In the bivariate case, the first factor is also an aggregate measure of industrial output. The second factor measures movements on the labour market and finally the last factor mixes employment variables and industrial production variables. These results seem coherent as in the univariate case we are not so much concerned with employment specifically but rather demand addressed to industrial companies. In this case industrial production, a healthy real estate market and global activity seem good predictors. In the second case in addition to industrial production, I want to forecast employment. Therefore in this case I mix industrial activity and employment components. This example shows us that even with a very large dataset I can give some interpretability to the factors as the grouping of the variables seem to represent some aspects of the economy.

Now to the model performances. In sample, and using all the components the fit of the univariate and bivariate model is at 0.99 for **x2**. This is due to the very large number of regressors that I have. The more interesting result concerns the forecasting performance. To do so, I do a one step ahead

rolling window forecast. Because I deal with time series data making cross validation difficult to implement, I use the BIC to tune the number of components in the model. This way I still mitigate the overfit and this tuning method provides comparable performance to regular cross-validation in this context like seen in Goulet Coulombe et al. 2022. The rolling window has a size of 200 observations. This means that to forecast one month in advance, I fit the data on the 200 past months, or roughly 16 years. I do not use the full sample as the data has structural breaks. For instance, if I predicted march 2020 with the full data this would mean that I forecast the covid period with data as old as the Bretton Woods period. In the mean time the Japanese economy has dramatically changed, so it would be like fitting the model on a different data generating process. Limiting it to 200 variables helps to balance the trade-off between increase in performance with increased number of data points for fitting, and decreased performance because of too old data being used for fitting the model.

In terms of Mean Squared Error I find the following performances. For the univariate model I have a MSE of 0.398, and for the bivariate model I have a MSE of 0.808 for **x2** and of 5.517 for **x106**. Here I see that moving from the univariate to the bivariate model slightly decreases the accuracy for **x2**. Nonetheless this decrease is moderate and shows how usefull the multivariate modelling can be using PLS regression.

Figure 1: 1 month ahead forecasts of the univariate model

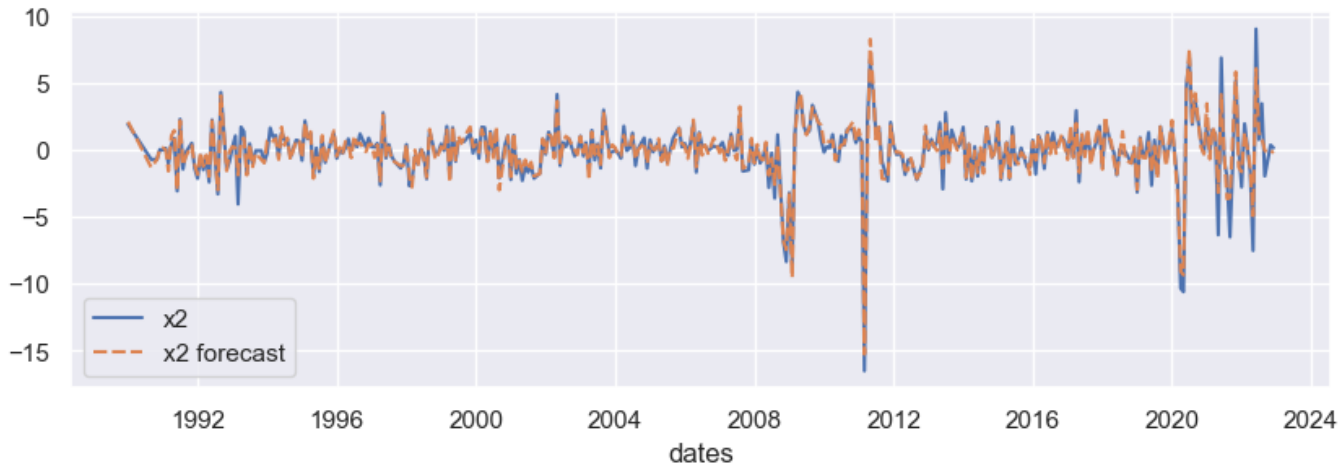
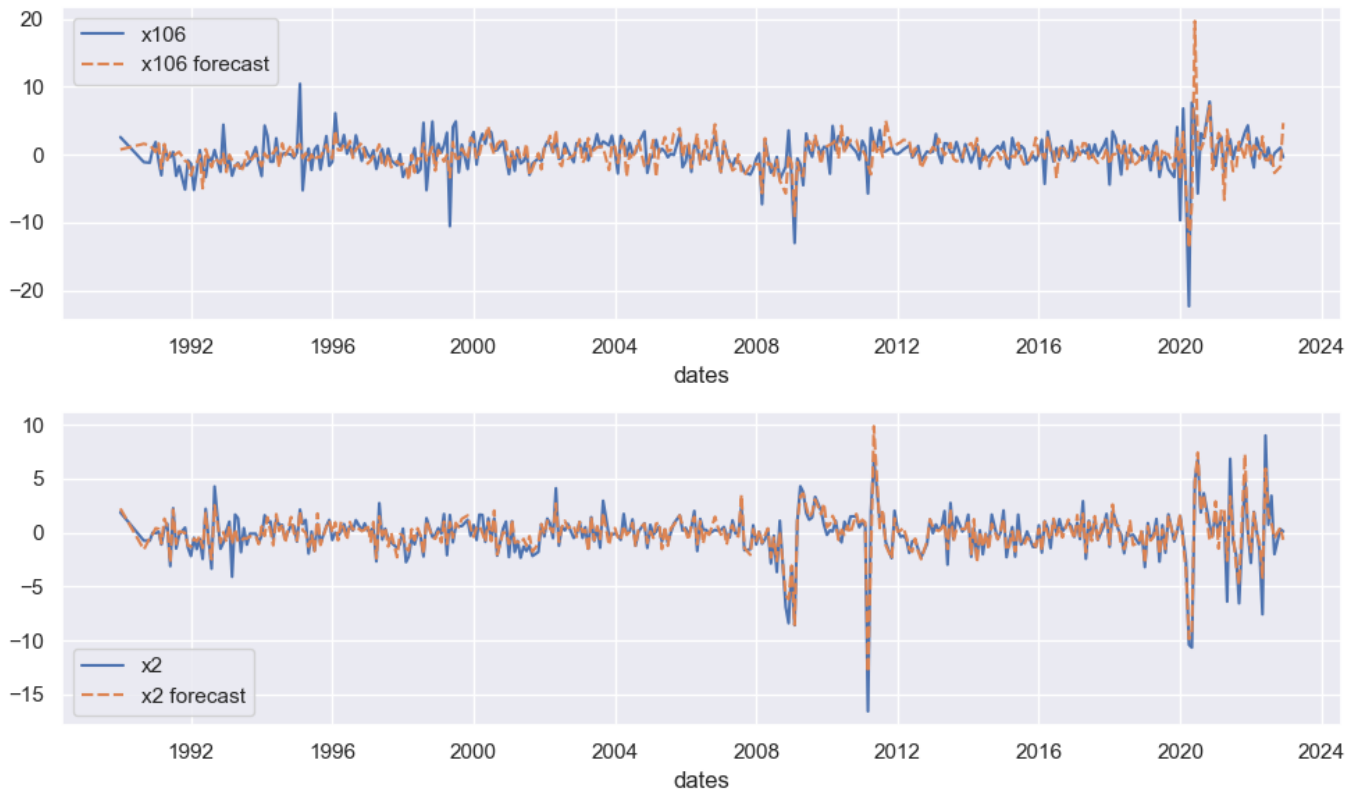


Figure 2: 1 month ahead forecasts of the bivariate model



5 Conclusion

First of all, before moving to the final conclusion, I compare the results with PLSRegression function of Sci-Kit learn. The results are available in the notebook. I do not find the same results as Sci-Kit learn. First of all I switched from the PLS algorithm presented in the lecture to the NIPALS algorithm used in Sci-Kit to match the results as close as possible. Even in this case the implementation of NIPALS in Sci-Kit remains different as for instance it doesn't initialize the loops for each components with the first column of the Y matrix but a specific one. Also it performs some data transformations that I was not able to pinpoint, this causes the forecasts of the Sci-Kit function to either be too volatile or not volatile enough. Also the very large number of predictors of the data set, and their high correlation probably doesn't help to get consistent results across repetitions or different methods. Nonetheless even with these differences, especially for the forecasts, the results and the one of the package mostly exhibit the same behaviour. In addition, even if factor loadings are not the same for the function and the package, the interpretation of the first 5 factors remains the exact same. The package creates similar, but not identical, factors that group the same kind of correlated variables.

References

Goulet Coulombe, Philippe et al. (2022). "How is machine learning useful for macroeconomic forecasting?" In: *Journal of Applied Econometrics* 37.5, pp. 920–964. DOI: https://doi.org/10.1002/jae.2910.

Maehashi, Kohei and Mototsugu Shintani (2020). "Macroeconomic forecasting using factor models and machine learning: an application to Japan". In: *Journal of the Japanese and International Economies* 58, p. 101104. ISSN: 0889-1583. DOI: https://doi.org/10.1016/j.jjie.2020.101104.