Big Data: Forecasting using a large number of predictors

Ahmad Tahmid

Università di Bologna

21st June 2024

Table of Contents

Brief Description of Data

- The dataset employed for the out-of-sample forecasting analysis is the same as the one used in Stock&Watson, De Mol.
- The panel includes real variables (sectoral industrial production, employment, and hours worked), nominal variables (consumer and producer price indices, wages, money aggregates), asset prices (stock prices and exchange rates), the yield curve, and surveys for a total of 131 variables.
- Sourced from St. Louis Fed.
- Required package"fbi" *devtools* :: *install_github*("*cykbennie/fbi*")
- Data period: 01/01/1959 to 01/12/2018, Monthly data.

Empirics: Transformation to Stationarity, Dealing with Missing Values, Standardization

- The series are transformed by taking logarithms and/or differencing to achieve stationarity.
- First differences of logarithms (growth rates) are used for real quantity variables, first differences for nominal interest rates, and second differences of logarithms for price series.
- Our variable of interest is CPIAUCSL (Consumer Price Index).
- For example CPI is transformed by the function

$$G_{it} = \Delta \ln \frac{z_{it}}{z_{it-12}} \times 100 \quad \text{monthly difference of yearly growth rate}$$

Empirics: Transformation to Stationarity, Dealing with Missing Values, Standardization

- Used the fredmd function which takes the syntax: fredmd(file = "", date_start = NULL, date_end = NULL, transform = TRUE)
- Used function *rm_outliers.fredmd* to control for outlier
- The dataset was split into a training set (80%) with 556 observations and a test set (20%) with 140 observations.

OLS: Ordinary Least Squares Regression

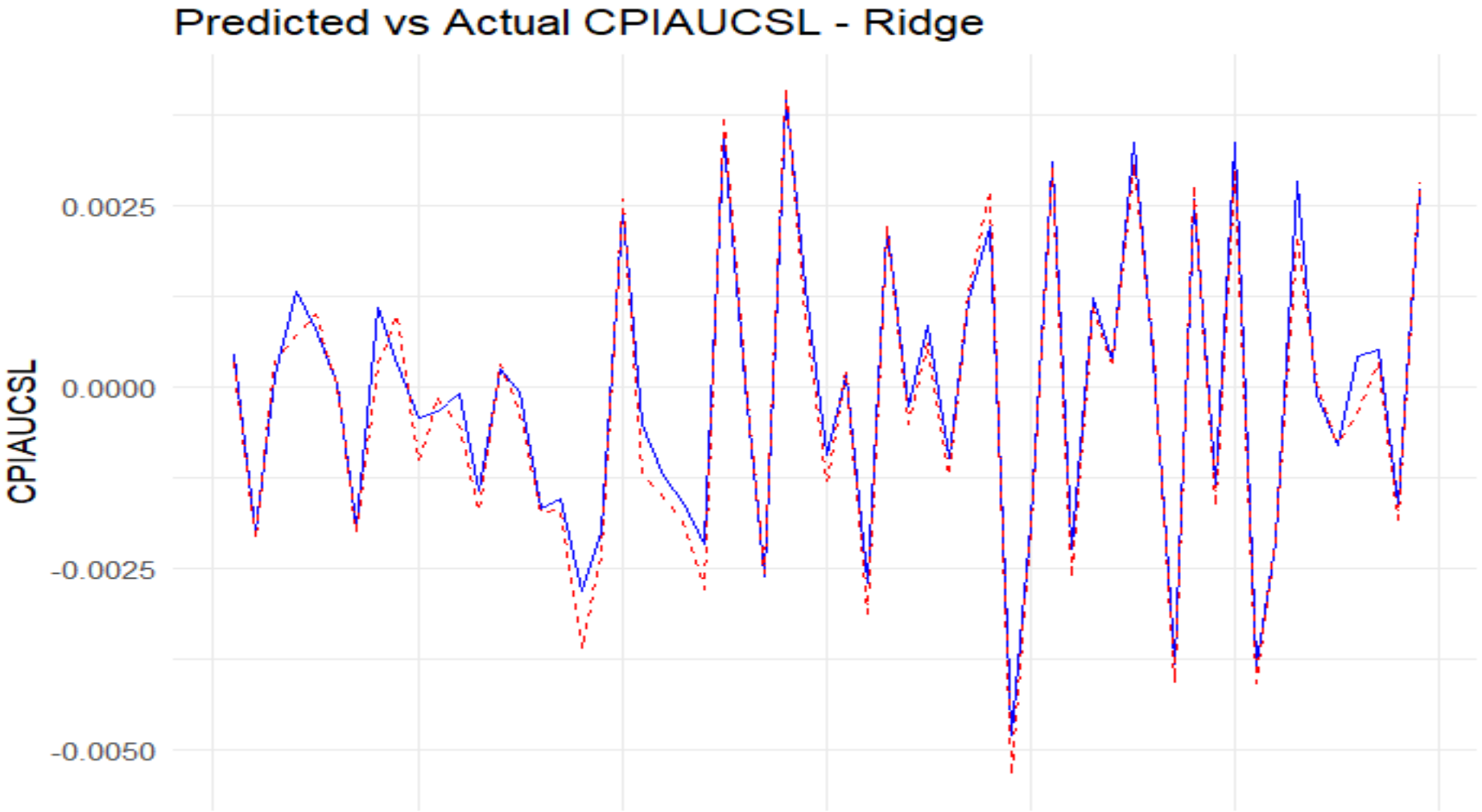
- Model: Autoregressive (AR) model.
- Lags: The optimal number of lags was determined using the Akaike Information Criterion (AIC), and 5 lags were selected.
- MSFE: The Mean Squared Forecast Error (MSFE) for the OLS model was approximately 3.908910e-06.
- To note:
  - Assumed data had no non-stationarity due to trends. No Dickey-Fuller tests were done.
  - Attempted to correct for non-stationarity by excluding data from COVID-19 years.

Ridge Regression

- Penalty Setting: Cross-validation was used to determine the optimal penalty.
- The optimal λ minimizes the cross-validated mean squared error.
- **Penalty Setting:** Cross-validation was used to determine the optimal penalty.
- **A design matrix** `X_train` is created with lagged values as predictors. The response variable `y_train` is the CPIAUCSL series.
- `cv.glmnet`: Performs 10-fold cross-validation to find the optimal value of λ (regularization parameter) for the Ridge regression model. The `alpha = 0`.
- **Optimal λ:** The λ that minimizes the cross-validated mean squared error is selected.
  - The choice of λ is a trade-off between bias and variance:
    - **Small λ:** Results in a model with lower bias but higher variance, potentially leading to overfitting.
    - **Large λ:** Results in a model with higher bias but lower variance, potentially leading to underfitting.

Ridge Regression

- MSFE: The MSFE for the Ridge Regression model was approximately 1.260912e-07(best of all!)



Principal Component Regression (PCR)

- Did a scree plot: Approximately 8 principal components explain more than 95% of the variance.

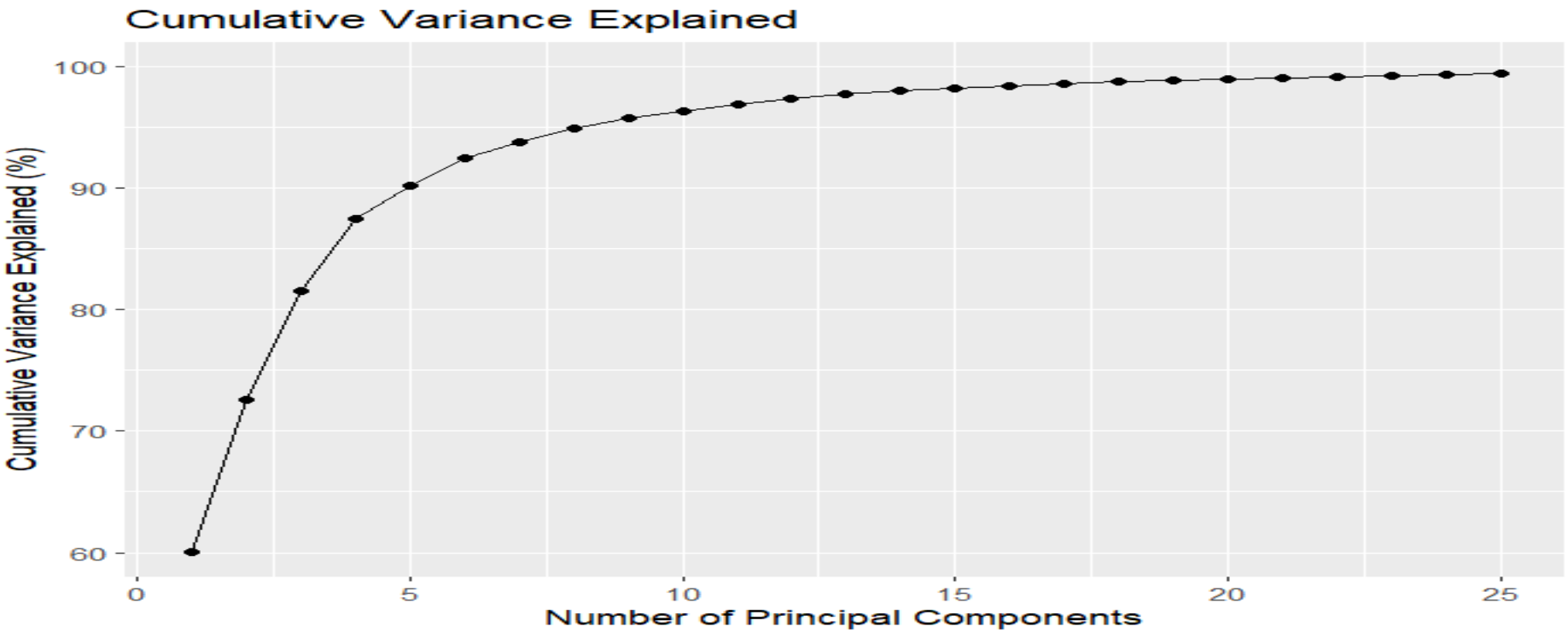


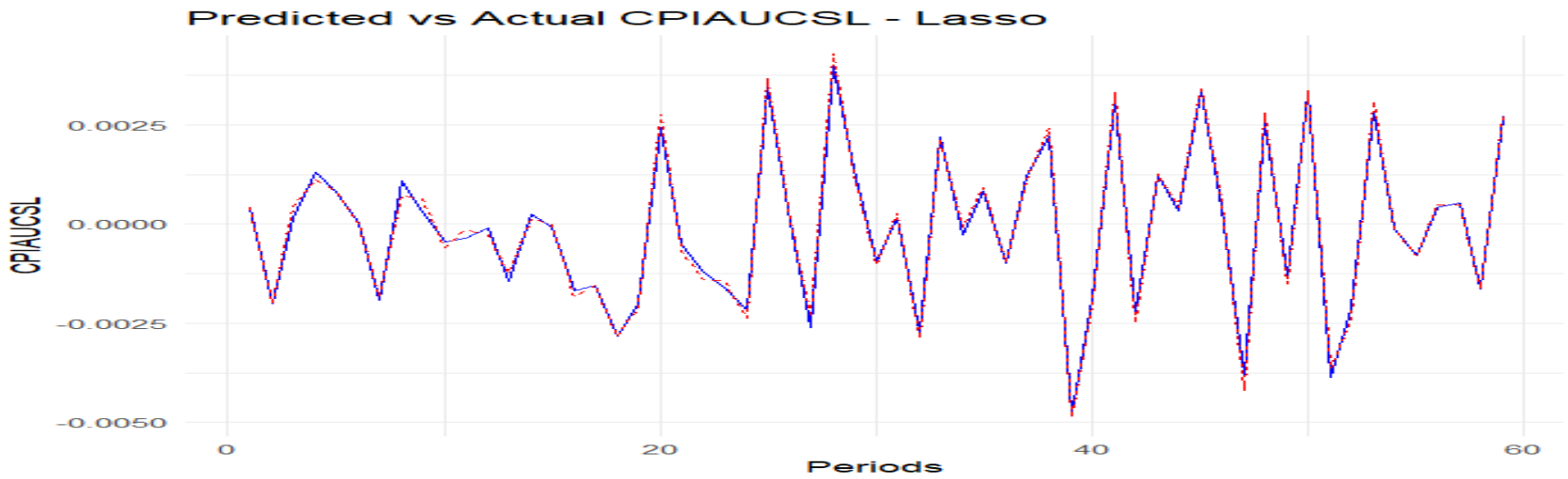
Figure: *I have limited the x axis to 25 for better visibility*

Principal Component Regression (PCR)

- The PCR model with 8 components resulted in an MSFE of approximately 1.641886e-07.
- **Notice** how the MSFE for PCR is larger. The PCR reduces dimensionality by focusing on the components that explain the most variance in the predictors. However, these components might not be the best at predicting the response variable (CPIAUCSL in this case). If the principal components that explain the most variance in the predictors don't correlate strongly with the response variable, PCR might not perform as well as Ridge Regression *I am not sure professor...just theorizing*

Lasso Regression

- Penalty Setting: Cross-validation was used to determine the optimal penalty.
- MSFE: The MSFE for the Lasso Regression model was approximately 3.016336e-08.
- Tends to overfit due to(??):
  - Too many variables/features.
  - High correlation among predictors.



Summary of Results and Conclusion

- Using the package"xtable" to export my sumamry results from R-studio

.

Method	MSFE
OLS (Autoregressive)	3.908910e-06
Ridge Regression	1.260912e-07
Lasso Regression	3.016336e-08
Principal Component Regression	1.641886e-07

Table: MSFE for Different Forecasting Methods

- Best performance for RIDGE followed by PCR followed by LASSO followed by OLS

Summary of Results and Conclusion (Continued)

- If multicollinearity is present, Ridge Regression might be the best option.
- If only a few predictors are important, Lasso might be superior.
- If the data is high-dimensional with many predictors, PCR could be the best approach.
- Based on the general characteristics of our data the best performances from $Ridge > PCR > LASSO > OLS$
- Consistent with our findings.

—x— Thank you!