Big Data Applications
Final Report

Authors: Luca Teresa; Dodon Alessandro; Marabini Giacomo.

The Data

The dataset we used is FRED_MD, a large monthly macroeconomic database. In the panel are included real variables divided in 8 groups, for instance Output and income (20 variables), Labor Market (32), Housing, Consumption (10), orders and inventories (14), Money and credit (14), Interest and exchange rates (22), Prices (21) and at the end Stock market (5). The total number of variables is 135, monthly reported from January 1st, 1959, to February 1st, 2024.

Data Cleaning and Transformation

After plotting the variables, it is evident that most of them suffer from non-stationarity and autocorrelation issues. For instance, we include the plot of Inflation (CPIULFSL) (image 1). Adjustments are necessary and we exploit the fredm function to do so. The transformations' results can be visualized through the plot of Inflation made stationary, (see image 2). To further observe if the above mentioned dynamics have been amended, it is useful to observe the autocorrelation plot (image 3).

The dataset was further manually cleaned, missing values not handled by the fredm function were identified and removed as follows. Firstly, we detected five variables[1] with extensive missing values, leading to the removal of these predictors. Additionally, we observed many missing values in the year 1959 thus we removed the year, such that our analysis starts from January 1960. The same applied to the last observation of the year 2024. This left us with a few scattered NAs across the dataset, which we addressed by applying a simple moving average algorithm.

Train and Test Samples

For the choice of splitting the sample, we have decided to assign the training sample to the period January 1st, 1960, to December 1st, 1999 (two thirds of the sample). The test sample is the remaining third of the entire sample, from January 1st, 2000, to December 1st, 2019.

Forecasting Methods

For the purpose of our analysis, we note that in the application of the models we have taken into account the data up to December 2019, in order to avoid a Covid-19 bias. Additionally, for each model that follows, a loop was constructed to: standardize the data, apply the regression model, predict one observation and de-standardize it, add one observation from the test data to the train data and repeat at each iteration. Once predictions were computed for the length of the time period we compared them with actual values.

Let's see the models more in depth.

[1] New Orders For Consumer Goods, New Order For Nondefense Capital Goods, Trade Weighted U.S. Dollar Index, Consumer Sentiment Index, VIX.

Ordinary Least Squares

In order to conduct an appropriate analysis using the OLS method we have decided to compute two different regressions: in the first one we have computed the autoregressive model using the observed variable (CPIULFSL) as predictor, while in the second part we have regressed using all the variables.
Comparing these two regressions, it's noticeable that the first model (with only CPIULFSL as predictor) is much more precise in predicting than the second model (the one with all 121 variables). It is even more clear when looking at the errors of these two models, especially the MSEs, presented in the table below.

Ridge

The following step regards the Ridge regression with the aim of minimizing the sum of residual squares using a term of penalization λ. This term is extremely useful in order to reduce the coefficients' weight in the regression, and therefore to limit the model's complexity. Thus, we set λ close to zero, then close to infinity to observe its behavior and in the end we computed the optimal value being $\lambda = \frac{p}{\sqrt{T_1}}$. The first two values deliver opposite results, namely overfitting for λ close to zero and underfitting for $\lambda = \frac{p}{\sqrt{T_1}}$, while it renders very well with $\lambda^{optimal}$.

Lasso

For the sake of this project, we wanted to observe the behavior of Lasso throughout different levels of penalty λ. Hence, we run the model with the following values: λ close to zero (0.0001), $\lambda = 0.01$, $\lambda = 0.10$, $\lambda = 1$ and so on up to λ being close to infinity (10^6). Lastly, we also applied Ridge optimal value $\lambda = \frac{p}{\sqrt{T_1}}$. We observed that the Lasso regression model transitioned from being approximately equivalent to OLS when $\lambda = 0.0001$, resulting in severe overfitting, to being completely underfitting from $\lambda = 1$ to $\lambda = 10^6$, as the algorithm effectively eliminated all variables, selecting virtually none. An optimal solution that balanced both a good fit and a low MSE was achieved with $\lambda = 0.01$ (see image 9).

Principal Component Regression

The primary step in Principal Component Analysis has been the evaluation of the covariance and the correlation of the variables considered. The two plotted matrices, present in the appendix, allowed us to visualize the relationship among the variables.
In addition, we have also computed eigenvalues and eigenvectors for both two matrices, in order to establish the optimal value of components for the application of the model. From the graph of eigenvalues (image 5) can be noticed the path of distribution and the variance explained by each component. For the purpose of our analysis we have decided to take in consideration the first five principal components of the distribution, because, as it is clear looking at the graph, going from the sixth point forward, the difference from each component goes close to zero.
Furthermore we also computed three more PCRs using respectively 1, 2, 5 (see image 10) and finally all principal components. In the case of the one component analysis, the variance was around 15%, in the PCRs with 2 components the explained variance was around 22% and for 5 components the variance was less than 50%.
All the resulting MSEs are listed in the table below.
We notice that using all components results in an MSE equivalent to OLS with all predictors; in fact, this has been computed to assure the correct functioning of our PCA.

Results

Mean Squared Errors		
OLS	With one predictor or AR(1)	1.351572e-05
	With all predictors	1.666667e-05
Ridge	$\lambda = 0.0001$	3.889971e-05
	$\lambda = 10^6$	1.312739e-05
	$\lambda = \frac{p}{\sqrt{T_1}}$	1.264374e-05
Lasso	$\lambda = 0.0001$	3.873223e-05
	$\lambda = 0.01$	1.173288e-05
	$\lambda = 0.10$	1.198938e-05
	$\lambda = 1$	1.312739e-05
	$\lambda = 10^6$	1.312739e-05
	$\lambda = \frac{p}{\sqrt{T_1}}$	1.312739e-05
PCA	With 1 component	1.312878e-05
	With 2 components	1.328557e-05
	With 5 components	1.338589e-05
	With all components	1.666667e-05

From the results obtained, we can make the following observations.

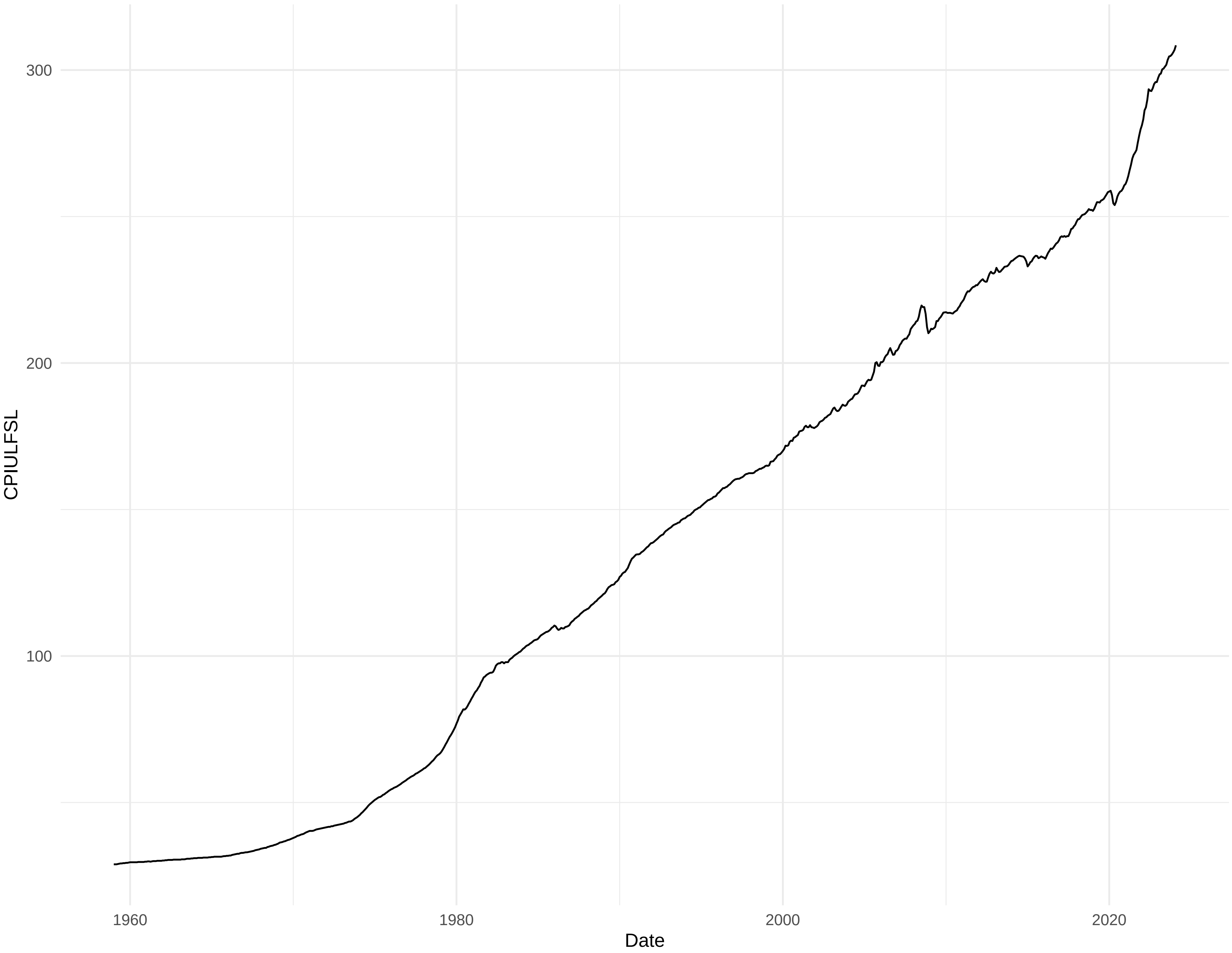
Regarding the AR(1) model, it results in a slightly lower MSE compared to using all predictors, suggesting that a simpler model can sometimes be more effective, likely due to less overfitting and multicollinearity.

Instead, for what concerns Ridge, with lambda close to infinity, the MSE is very low, but the model is not at all representative and it is the optimal value of lambda that provides the best MSE and predictions. This demonstrates the importance of selecting an appropriate penalty parameter to achieve a balanced model.

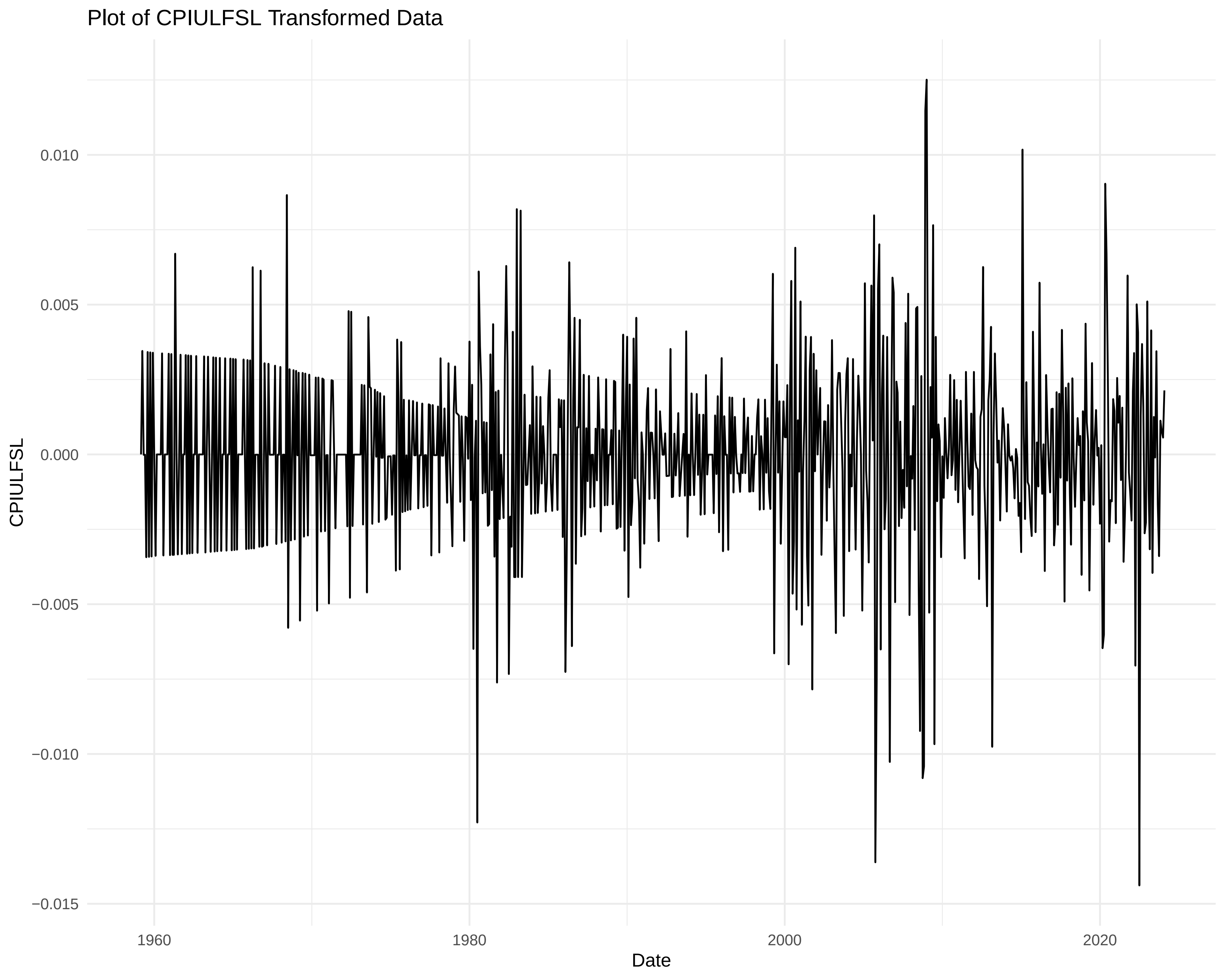
In Lasso regression the results are somewhat similar to Ridge. However, with lambda close to zero, the MSE is significantly higher. Thus even though graphically it seems an accurate specification, we can clearly infer that it suffers from overfitting.

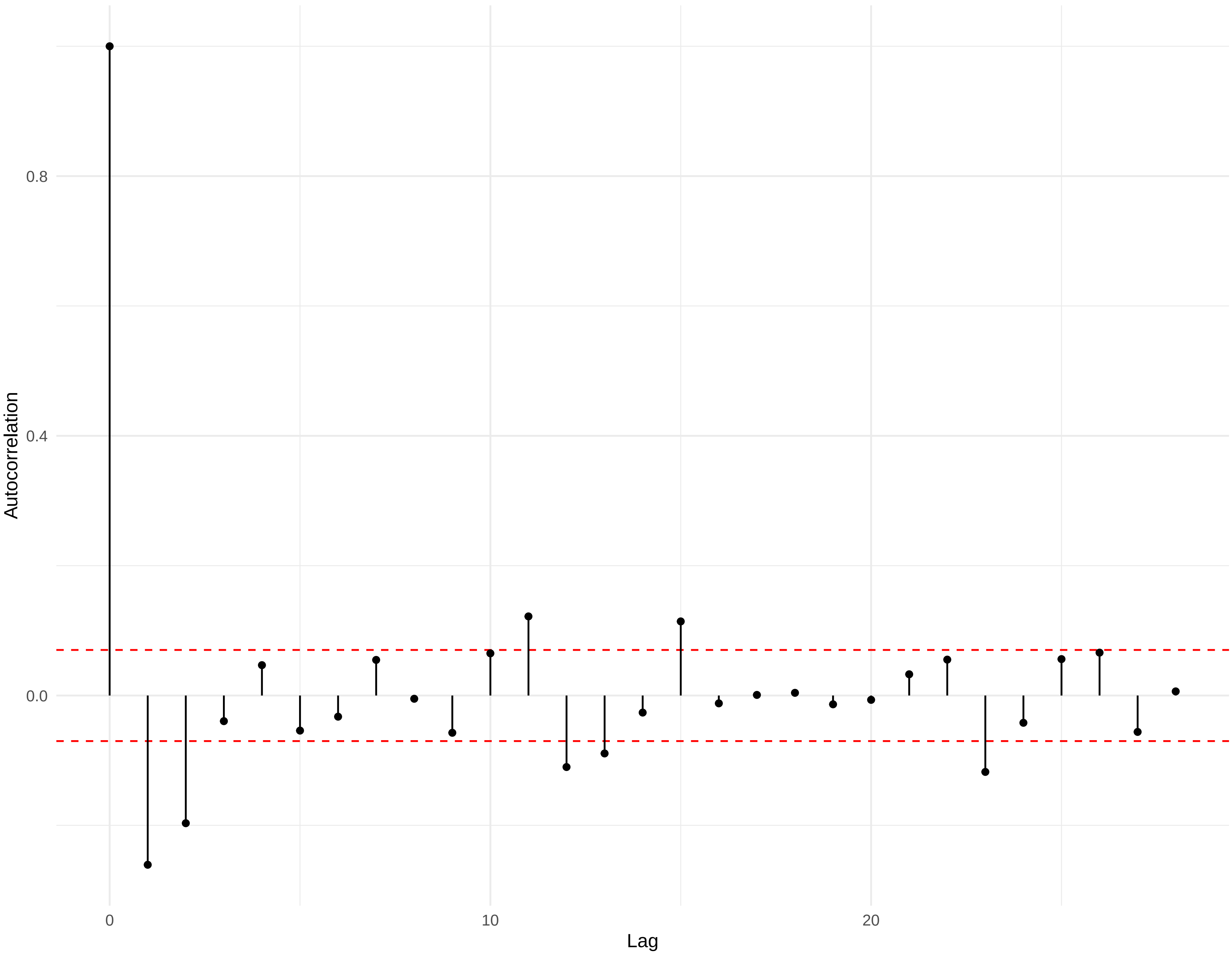
Exploiting Principal Component Regression and using a single component offers a very low MSE, but also a complete underfitting model. The same could be said for PCR with 2 and also 5 principal components, likely due to the small portion of variance explained by them.

Overall, the Ridge regression model with the optimal value of lambda, as well as the Lasso with a lambda set to 0.01 produced the best MSE and a good fit to the actual values. We believe this is because of the nature of these methods, as introducing regularization improves the bias-variance trade-off and results in more robust and generalizable models.

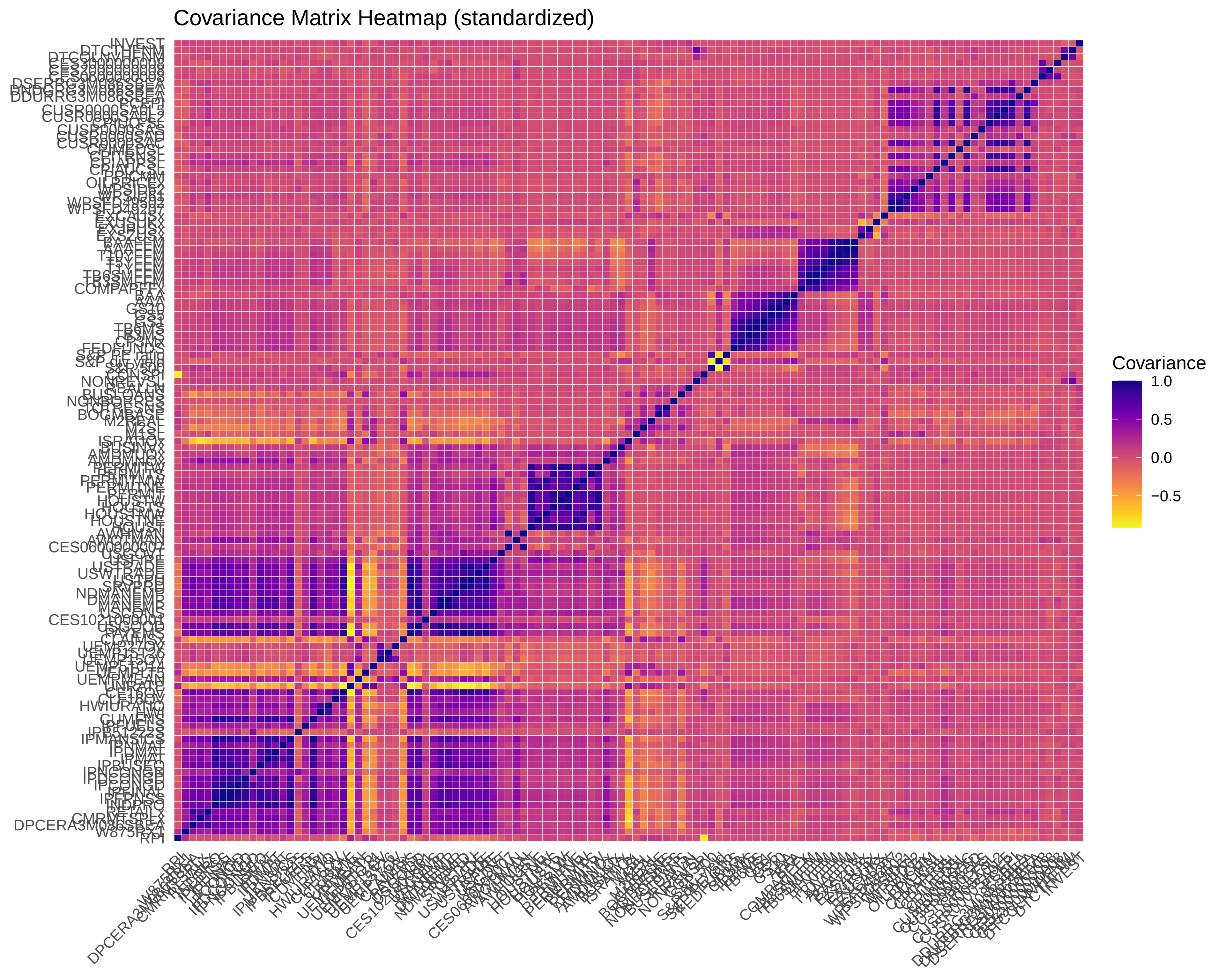


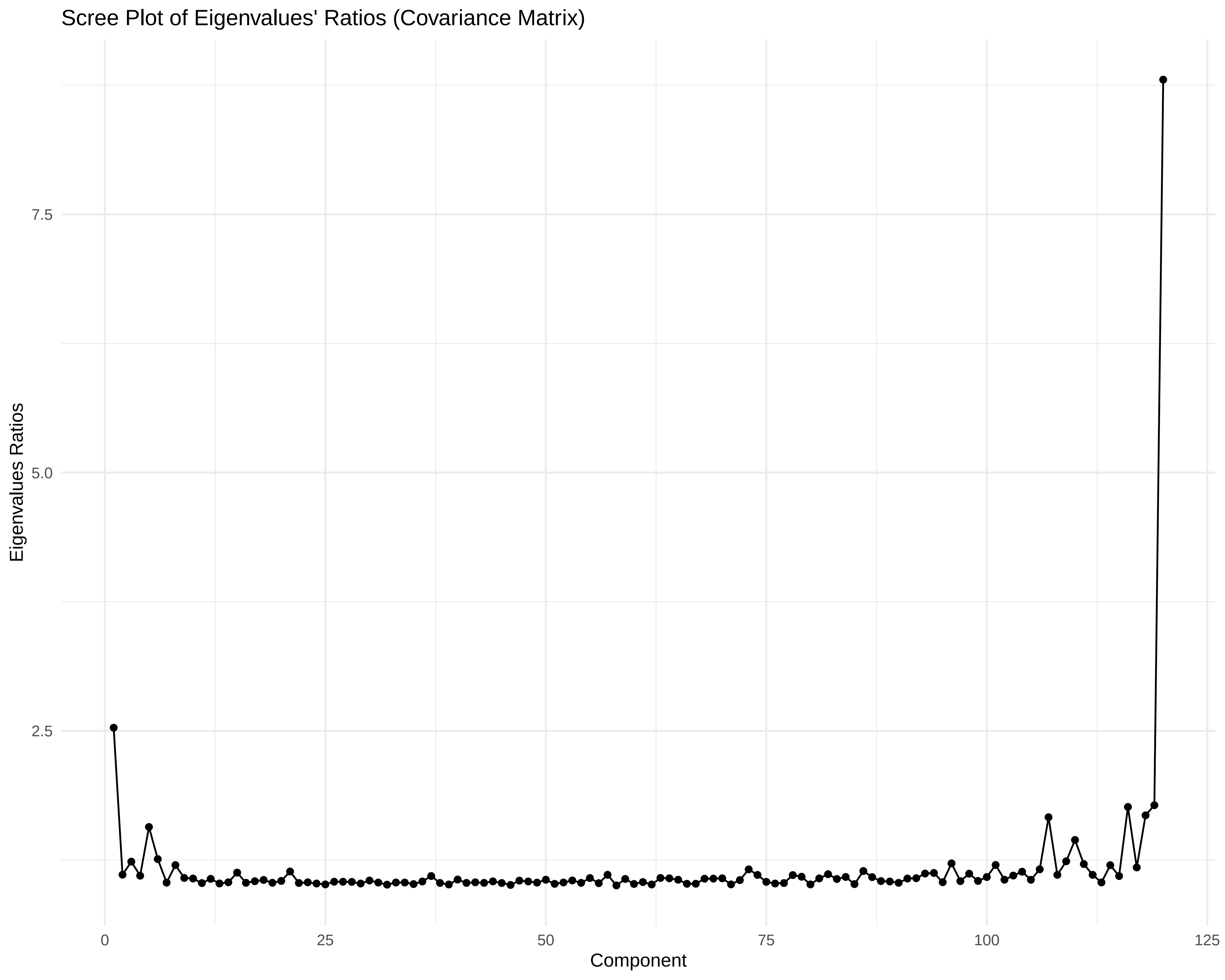
Plot of CPIULFSL Original Data

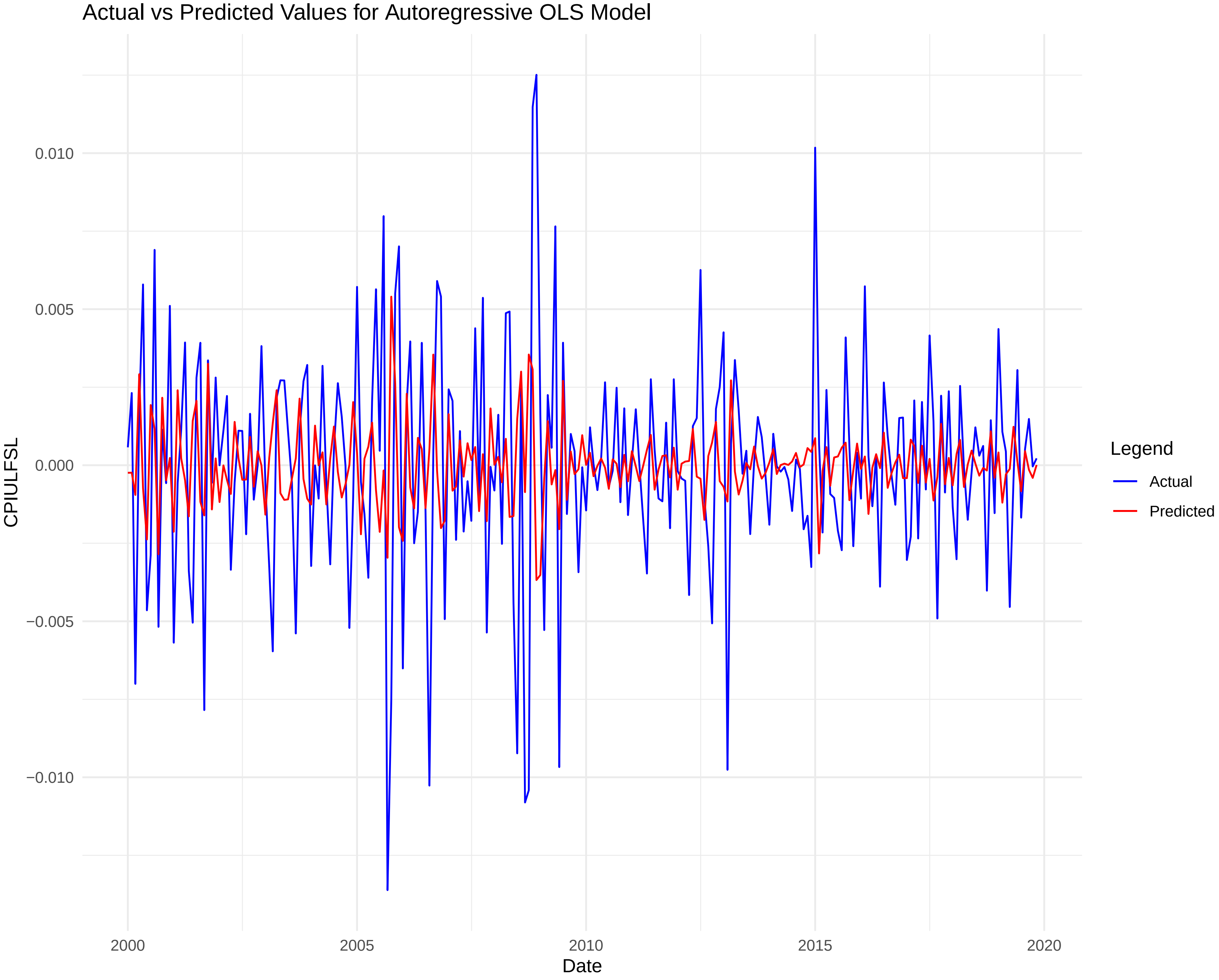

Plot of CPIULFSL Transformed Data
CPIULFSL
0.010
0.005
0.000
−0.005
−0.010
−0.015
Date
1960
1980
2000
2020

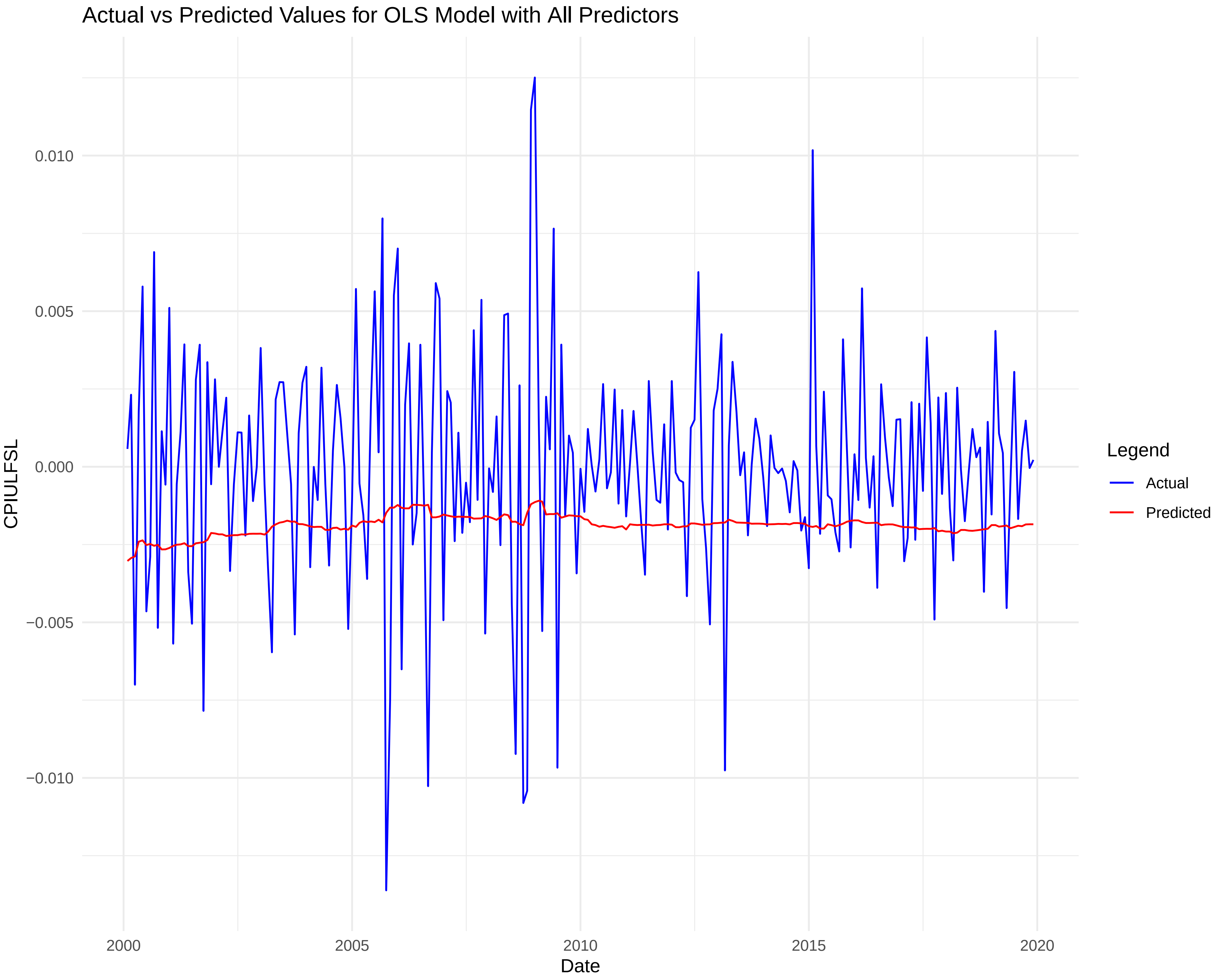


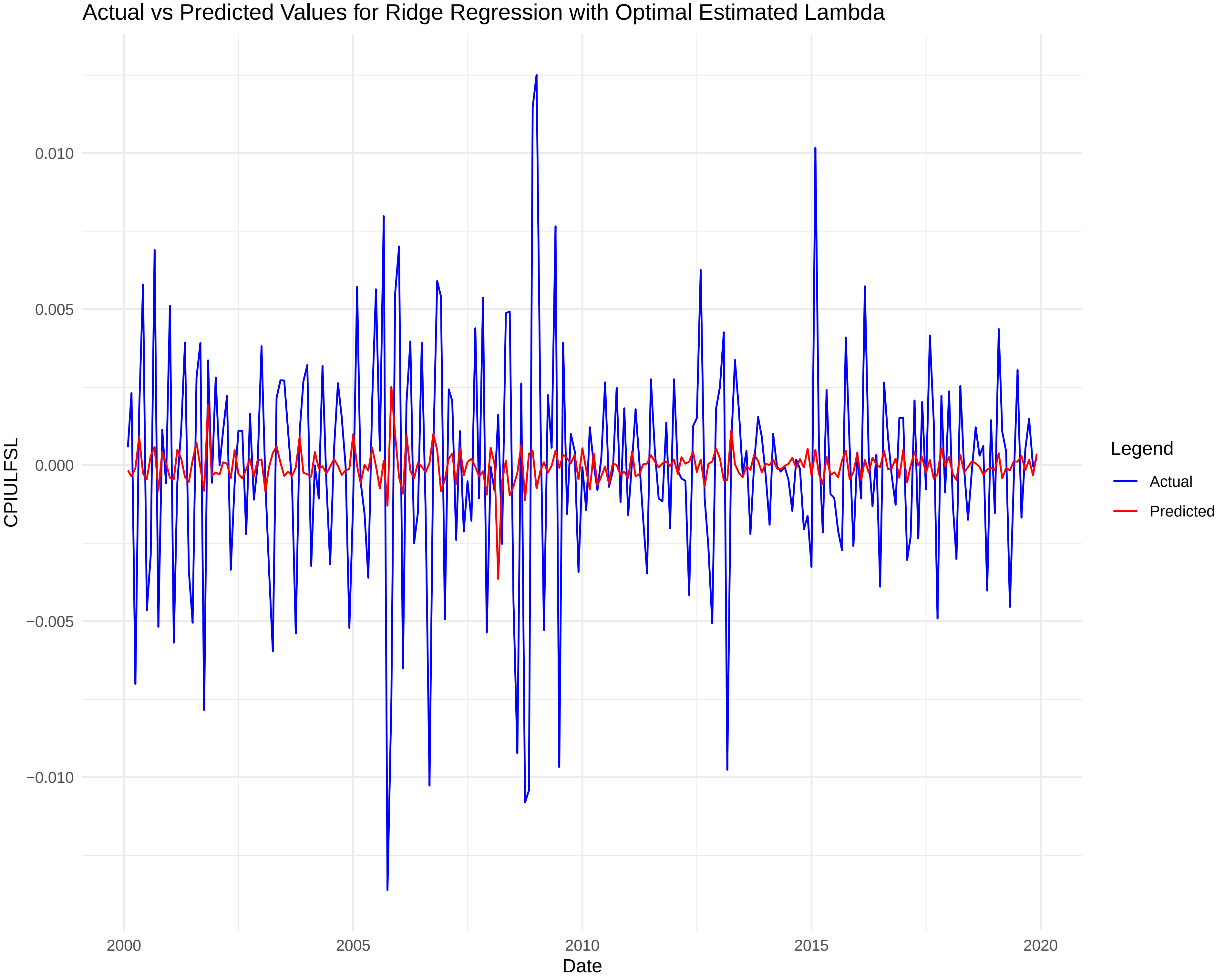
Autocorrelation of CPIULFSL Transformed Data

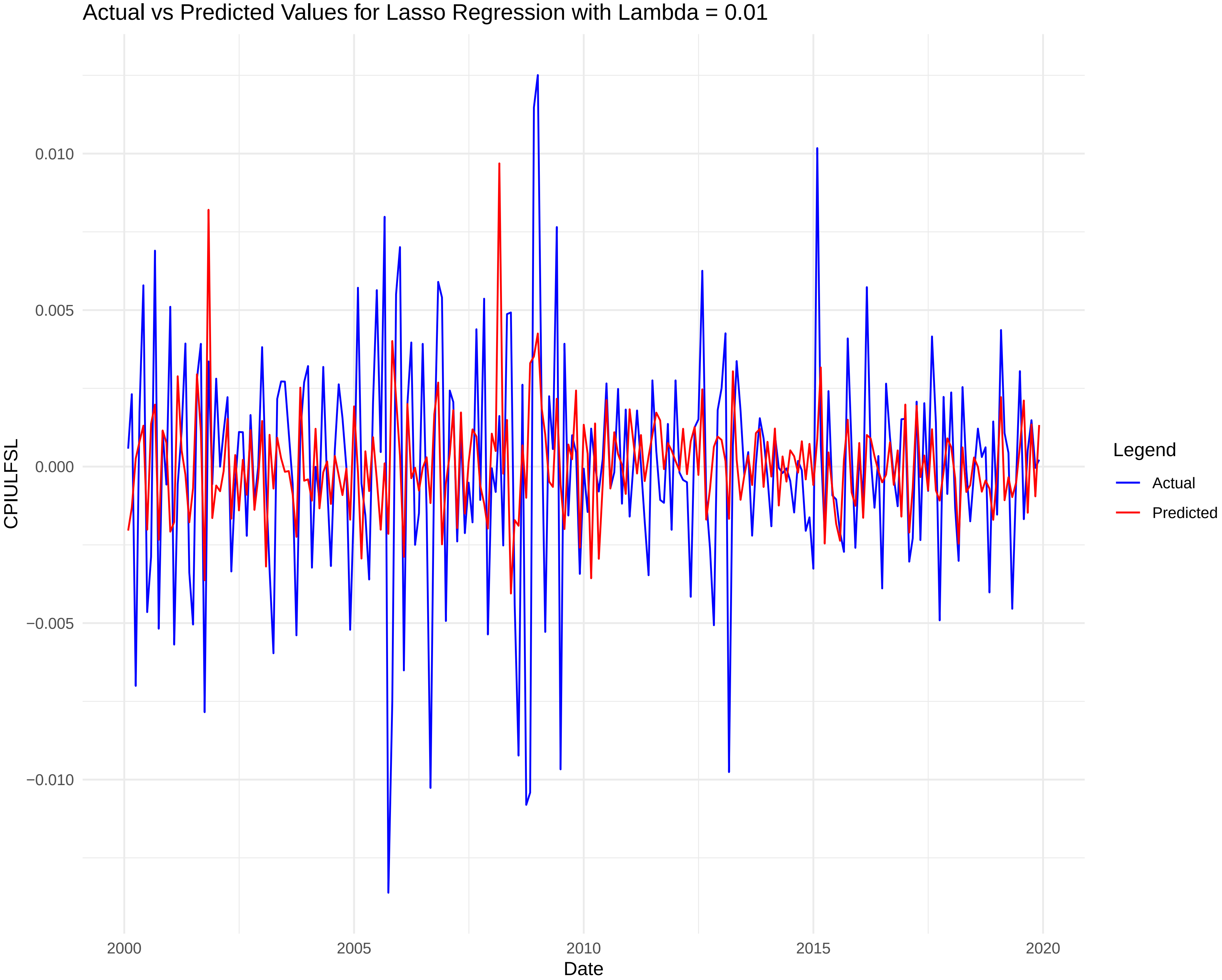

Covariance Matrix Heatmap (standardized)
Covariance
1.0
0.5
0.0
−0.5
INVEST
DTCTHFNM
DTCOLNVHFNM
CES3000000008
CES2000000008
CES0600000008
DSERRG3M086SBEA
DNDGRG3M086SBEA
DDURRG3M086SBEA
PCEPI
CUSR0000SA0L5
CUSR0000SA0L2
CPIULFSL
CUSR0000SAS
CUSR0000SAD
CUSR0000SAC
CPIMEDSL
CPITRNSL
CPIAPPSL
CPIAUCSL
PPICMM
OILPRICEx
WPSID62
WPSID61
WPSFD49502
WPSFD49207
EXCAUSx
EXUSUKx
EXJPUSx
EXSZUSx
BAAFFM
AAAFFM
T10YFFM
T5YFFM
T1YFFM
TB6SMFFM
TB3SMFFM
COMPAPFFx
BAA
AAA
GS10
GS5
GS1
TB6MS
TB3MS
CP3Mx
FEDFUNDS
S&P PE ratio
S&P div yield
S&P: indust
S&P 500
CONSPI
NONREVSL
REALLN
BUSLOANS
NONBORRES
TOTRESNS
BOGMBASE
M2REAL
M2SL
M1SL
ISRATIOx
BUSINVx
AMDMUOx
AMDMNOx
PERMITW
PERMITS
PERMITMW
PERMITNE
PERMIT
HOUSTW
HOUSTS
HOUSTMW
HOUSTNE
HOUST
AWHMAN
AWOTMAN
CES0600000007
USGOVT
USFIRE
USTRADE
USWTRADE
USTPU
SRVPRD
NDMANEMP
DMANEMP
MANEMP
USCONS
CES1021000001
USGOOD
PAYEMS
CLAIMSx
UEMP27OV
UEMP15T26
UEMP15OV
UEMP5TO14
UEMPLT5
UEMPMEAN
UNRATE
CE16OV
CLF16OV
HWIURATIO
HWI
CUMFNS
IPFUELS
IPB51222S
IPMANSICS
IPNMAT
IPDMAT
IPMAT
IPBUSEQ
IPNCONGD
IPDCONGD
IPCONGD
IPFINAL
IPFPNSS
INDPRO
RETAILx
CMRMTSPLx
DPCERA3M086SBEA
W875RX1
RPI

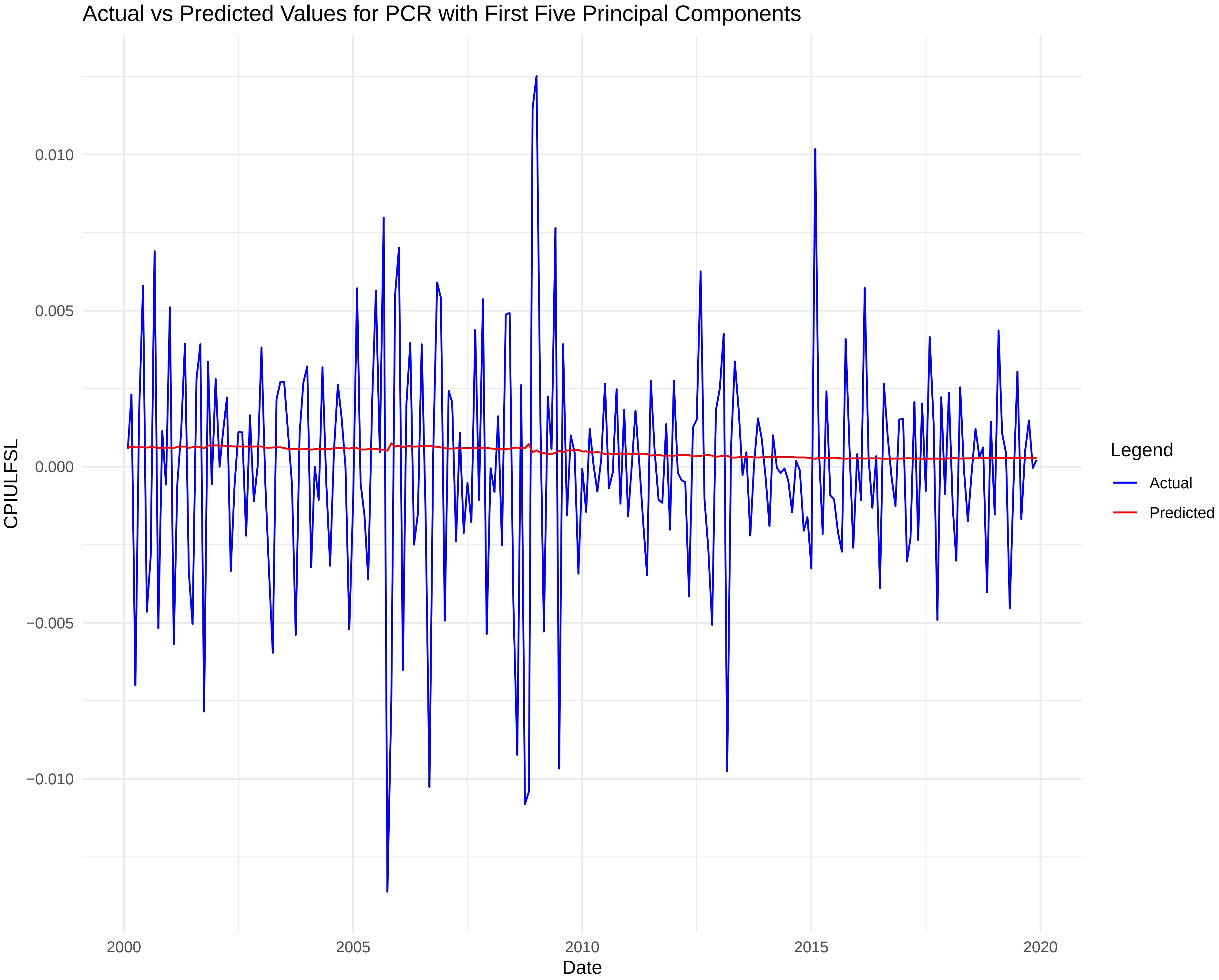

Scree Plot of Eigenvalues' Ratios (Covariance Matrix)
Eigenvalues Ratios
7.5
5.0
2.5
0
25
50
75
100
125
Component


Actual vs Predicted Values for Autoregressive OLS Model
0.010
0.005
0.000
−0.005
−0.010
CPIULFSL
2000
2005
2010
2015
2020
Date
Legend
Actual
Predicted


Actual vs Predicted Values for OLS Model with All Predictors
CPIULFSL
0.010
0.005
0.000
−0.005
−0.010
2000
2005
2010
2015
2020
Date
Legend
Actual
Predicted


Actual vs Predicted Values for Ridge Regression with Optimal Estimated Lambda
CPIULFSL
0.010
0.005
0.000
−0.005
−0.010
2000
2005
2010
2015
2020
Date
Legend
Actual
Predicted


Actual vs Predicted Values for Lasso Regression with Lambda = 0.01
CPIULFSL
0.010
0.005
0.000
−0.005
−0.010
2000
2005
2010
2015
2020
Date
Legend
Actual
Predicted


Actual vs Predicted Values for PCR with First Five Principal Components
CPIULFSL
0.010
0.005
0.000
−0.005
−0.010
Date
2000
2005
2010
2015
2020
Legend
Actual
Predicted