Machine Learning in Finance

Overview

- Big Data Issues
- Forecasting Methods and Evaluation, Part I
- Forecasting Methods and Evaluation, Part II
- Forecasting Methods and Evaluation, Part III
- Regularization with Elastic Net
- Information Criteria, LASSO and Cross Validation
- Example with Monte Carlo Simulation
- Forecasting S&P 500 Weekly Returns
- Forecasting Daily Volatility of the S&P 500 Index

FORDHAM
THE JESUIT UNIVERSITY OF NEW YORK

Gabelli School of Business

Big Data Issues

Big Data Issues

- When we apply models to big wide data sets, we have the problem of too many parameters
- As noted, too many parameters makes things difficult on many dimensions.
- We look for parsimony in models, that is we want to start with the simplest models and only gradually increase complexity
- We also know that danger of overfitting a model: models with lots of parameters fit the in-sample data sets well, but fen fail miserably with out of sample tests.
- Basic message: if a model gives too good a fit, that is bad news.
- Do not be an R^2 maximizer for the regression mode:

$$y_t = \sum \beta_k x_{k,t} + \epsilon_t \tag{1}$$

$$\epsilon_t \sim N(0, \sigma^2) \tag{2}$$

- We need to penalize ourselves for too many parameters, but how?

Big Data Issues

- Older ways to evaluate and rank competing models with different numbers of parameters are the Akaike, Schwartz and Hannan-Quinn Information Criteria. All of these statistics are called "information criteria".
- There are also known as regularization criteria. Idea is level the playing field for competing models.
- The last one handicaps or "punishes" the performance of a model for the number of parameters, k it uses:

$$hqif = \left[\ln\left(\sum_{t=1}^{T}\frac{(y_t - \widehat{y}_t)^2}{T}\right)\right] + \frac{k\{\ln[\ln(T)]\}}{T} \tag{3}$$

- The criterion is simply to choose the model with the lowest value.

Big Data Issues

- Note that the statistic punishes a given model by a factor of $k\{\ln[\ln(T)]\}/T$, the logarithm of the logarithm of the number of observations, T, multiplied by the number of parameters, k, divided by T.
- The Akaike criterion replaces the second term on the right-hand side of equation (2.40) with the variable $2k/T$, whereas the Schwartz criterion replaces the same term with the value $k[\ln(T)]/T$. Most financial folks work with the Hannan-Quinn statistic rather than the Akaike or Schwartz criteria, on the grounds that *virtu stat in media.*
- The Hannan-Quinn statistic usually punishes a model with too many parameters more than the Akaike statistic, but not as severely as the Schwartz statistic.

FORDHAM
THE JESUIT UNIVERSITY OF NEW YORK

Gabelli School of Business

Forecasting Methods and Evaluation, Part I

Forecasting Methods and Evaluation, Part I

- As noted, too many parameters deliver good in-sample fits but generally fail miserably with out-of-sample criteria.
- We usually measure out of sample performance with a given forecast interval, fit the exogenous or per-determined variables to the data with the old in-sample coefficients and see how the old coefficients work with the new data.
- Good in-sample performance, judged by the R^2 or the Hannan-Quinn statistics, may simply mean that a model is picking up peculiar or idiosyncratic aspects of a particular sample, or "over-fitting" the sample, but the model may not fit the wider population very well.
- To evaluate the out-of-sample performance of a model, we begin by dividing the data into an "in-sample" estimation or "training" set, for obtaining the coefficients, and an "out-of-sample" or "test" set.
- With the latter set of data, we plug in the coefficients obtained from the "training" set to see how well they perform with the new data set, which had no role in the calculating of the coefficient estimates.
- For time series forecasting, the out-of-sample performance can be calculated in two ways. One is simply to withhold a given percentage of the data for the test , usually the last two years of observations. The errors come from one set of coefficients, based on the fixed "training set", and one fixed "test set" of several observations.

Forecasting Methods and Evaluation, Part I

- An alternative to a once-and-for-all division of the data into "training" and "test" sets is the recursive methodology, which Stock (1999) describes as a series of "simulated real time forecasting experiments".
- It is also known as estimation with a "moving" or "sliding" window. In this case, period-by-period forecasts of variable y at horizon h, $\widehat{y}_{t+h}$, are conditional only on data up to time t. Thus, with a given data set, we may use the first half of the data, based on observations $\{1, ...t^*\}$ for the initial estimation, and obtain an initial forecast $\widehat{y}_{t^*+h}$.
- Then we re-estimate the model based on observations $\{1, ...t^* + 1\}$, and obtain a second forecast error, $\widehat{y}_{t^*+1+h}$. The process continues until the sample is covered.
- Needless to say, the many re-estimations of the model required by this approach can be "computationally. demanding" for nonlinear models. We call this type of recursive estimation an "expanding window".
- The sample size, of course, becomes larger as we move forward in time.

Forecasting Methods and Evaluation, Part I

- An alternative to the expanding window is the moving window.
- In this case, for the first forecast, we estimate with data observations $\{1, ...t^*\}$, and obtain the forecast $\widehat{y}_{t^*+h}$ at horizon h. We then incorporate the observation at $t^* + 1$, and re-estimate the coefficients with data observations $\{2, ...t^* + 1\}$, and not $\{1, ...t^* + 1\}$. The advantage of the moving window is that as data become more distant, in the past, we assume that it has little or no predictive relevance, so they are removed from the sample.
- The recursive methodology, as opposed to the once-and-for-all "split" of the sample, is clearly biased toward a linear model, since there is only one forecast error for each "training set".
- The linear regression coefficients adjust to and approximate, step-by-step, in a recursive manner, the underlying changes in the slope of the model, as they forecast out only one-step ahead.
- The appeal of the recursive linear estimation approach is that it reflects how econometricians do in fact operate. The coefficients of linear models are always being updated as new information becomes available, if for no other reason, linear estimates are very quick to obtain.
- It is hard to conceive of any organization using information a few years old to estimate coefficients for making decisions in the present.

FORDHAM
THE JESUIT UNIVERSITY OF NEW YORK

Gabelli School of Business

Forecasting Methods and Evaluation, Part II

Forecasting Methods and Evaluation, Part II

- The most commonly used statistic for evaluating out-of-sample fit is the Root Mean Squared Error (rmsq) statistic:

$$rmsq = \sqrt{\frac{\sum_{\tau=1}^{\tau^*}(y_\tau - \widehat{y}_\tau)^2}{\tau^*}} \tag{4}$$

where τ^* is the number of observations in the test set and $\{\widehat{y}_\tau\}$ are the predicted values of $\{y_\tau\}$.

- The out-of-sample predictions are calculated by using the input variables in the test set, $\{x_\tau\}$ with the parameters estimated with the in-sample data.
- We should select the model with the lowest root mean squared error statistic.
- However, how can we determine if the out-of-sample fit of one model is "significantly" better, that is, lower, than the out-of-sample fit of another model? One simple approach is to keep track of the out-of-sample points in which model A beats model B.

Forecasting Methods and Evaluation, Part II

- A more detailed solution to this problem comes from the work of Diebold and Mariano

Diebold-Mariano Procedure

Definition	*Operation*
Errors	$\{\widehat{\epsilon_\tau}\}, \{\widehat{\eta_\tau}\}$
Absolute differences	$z_\tau = \lvert\widehat{\eta_\tau}\rvert - \lvert\widehat{\epsilon_\tau}\rvert$
Mean	$\overline{z} = \frac{\sum_{\tau=1}^{\tau*} z_\tau}{\tau*}$
Covariogram	$c = [Cov(z_\tau, z_{\tau-p,\ }), Cov(z_\tau, z_{\tau,\ }), Cov(z_\tau, z_{\tau+p,\ })]$
Mean	$\overline{c} = \sum c/(p+1)$
DM statistic	$DM = \frac{\overline{z}}{\overline{c}} \sim N(0,1), H_0 : E(z_\tau) = 0$

Forecasting Methods and Evaluation, Part II

- As shown above, we first obtain the out-of-sample prediction errors of the "benchmark" model, given by $\{\epsilon_\tau\}$, as well as those of the competing model, $\{\eta_\tau\}$.
- Next, we compute the absolute values of these prediction errors, as well as the mean of the differences of these absolute values, z_τ.
- We then compute the covariogram for lag/lead length p, for the vector of the differences of the absolute values of the predictive errors. The parameter $p < \tau^*$, the length of the out-of-sample prediction errors.
- In the final step, we form a ratio of the means of the differences over the covariogram. The DM statistic is distributed as a standard normal distribution under the null hypothesis of no significant differences in the predictive accuracy of the two models.
- Thus, if the competing model's predictive errors are significantly lower than those of the benchmark model , the DM statistic should be below the critical value of -1.69 at the five percent critical level.

FORDHAM
THE JESUIT UNIVERSITY OF NEW YORK

Gabelli School of Business

Forecasting Methods and Evaluation, Part III

Forecasting Methods and Evaluation, Part III

- Harvey, Leybourne, and Newbold suggest a "size correction" to the DM statistic, which also allows "fat tails" in the distribution of the forecast errors. We call this modified Diebold-Mariano statistic the MDM statistic.
- It is obtained by multiplying the DM statistic by the correction factor CF, and it is asymptotically distributed as a Student's t with $\tau^* - 1$ degrees of freedom. The following equation system summarizes the calculation of the MDM test, with the parameter p representing the lag/lead length of the covariogram, and τ^* the length of the out-of-sample forecast set:

$$CF = \frac{\tau^* + 1 - 2p + p(1-p)/\tau^*}{\tau^*} \tag{5}$$

$$MDM = CF \cdot DM \sim t_{\tau^*-1}(0,1) \tag{6}$$

- Out-of-sample forecasts can also be evaluated by comparing the signs of the out-of-sample predictions with the true sample. In financial time series, this is particularly important if one is more concerned about the sign of stock return predictions rather than the exact value of the returns.
- After all, if the out-of-sample forecasts are correct and positive, this would be a signal to buy, and if they are negative, a signal to sell.

Forecasting Methods and Evaluation, Part III

Pesaran-Timmerman Directional Accuracy (DA) Test

Definition	*Operation*
Calculate out of sample predictions, m periods	$\widehat{y}_{n+j}, j = 1, \ldots m$
Compute Indicator for Correct Sign	$I_j = 1$ if $\widehat{y}_{n+j} \cdot y_{n+j} > 0, 0$ otherwise
Compute Success Ratio (SR)	$SR = \frac{1}{m} \sum_{j=1}^{m} I_j$
Compute Indicator for True Values	$I_j^{true} = 1$ if $y_{n+j} > 0, 0$ otherwise
Compute Indicator for Predicted Values	$I_j^{pred} = 1$ if $\widehat{y}_{n+j} > 0, 0$ otherwise
Compute Means P, $\widehat{P}$	$P = \frac{1}{m} \sum_{j=1}^{m} I_j^{true}, \widehat{P} = \frac{1}{m} \sum_{j=1}^{m} I_j^{pred}$
Compute Success Ratio under Independence (SRI)	$SRI = P \cdot \widehat{P} - (1 - P) \cdot (1 - \widehat{P})$
Compute Variance for SRI	$var(SRI) = \frac{1}{m}(2\widehat{P} - 1)^2 P(1 - P) + (2P - 1$ $+ \frac{4}{m} P \cdot \widehat{P}(1 - P)(1 - \widehat{P})]$
Compute Variance for SR	$var(SR) = \frac{1}{m} SRI(1 - SRI)$
Compute DA Statistic	$DA = \frac{SR - SRI}{\sqrt{var(SR) - var(SRI)}} \overset{a}{\sim} N(0, 1)$

Forecasting Methods and Evaluation, Part III

- One simple approach is to divide the initial data set into k subsets of approximately equal size. We then estimate the model k times, each time leaving out one of the subsets. We can compute a series of mean squared error measures on the basis of "forecasting" with the omitted subset. For k equal to the size of the initial data set, this method is called "leave out one"
- LeBaron proposes a more extensive bootstrap test, called the "0.632 bootstrap", originally due to Efron (1979) and described in Efron and Tibshirani
- The basic idea, according to LeBaron, is to estimate the original "in-sample" bias by repeatedly drawing new samples from the original sample, with replacement, and using the new samples as estimation sets, with the remaining data from the original sample, not appearing in the new estimation sets, as "clean" test or "out-of-sample" data sets.
- In each of the repeated draws, of course, we keep track of which data points are in the "estimation" set and which are in the "out-of-sample" data set.
- Depending on the draws in each repetition, the size of the out-of-sample data set will vary.

FORDHAM
THE JESUIT UNIVERSITY OF NEW YORK

Gabelli School of Business

Regularization with Elastic Net

Regularization with Elastic Net

- An alternative to the earlier Information Criteria is to use the **Elastic Net** for parameter reduction or regularization:

$$\beta_{Enet} = \overset{Min}{\beta} \left\{ \sum_{t=1}^{T} \left(y_t - \sum_{i} \beta_i x_{it} \right)^2 + \lambda \sum_{i=1}^{k} \left[(\alpha|\beta_i|) + (1-\alpha)\beta_i^2 \right] \right\} \tag{7}$$

- Not that the penalty terms applies to the size of the sum of squared coefficients or the sum of the absolute value of the coefficients, or some combination of both.
- The elastic net combines the LASSO and Ridge penalties through the **tuning parameters** or **hyperparameters** $\{\alpha, \lambda\}$. With $\alpha = 1, \lambda > 0$, it is a LASSO (Least Absolute Shrinkage Selection Operator), it is a Ridge estimator with $\alpha = 0$,$\lambda > 0$. With $\lambda = 0$,of course, there is no penalty for large numbers of parameters, and the estimates are equivalent to the least-squares.

Regularization with Elastic Net

- The OLS estimator, with no penalty for large numbers of parameters, would allow for large numbers of small, insignificant coefficients
- By making use of the Elastic Net, we are eliminating variables which have small absolute or squared values. Note we are not using a T-statistic.
- With this net, the choice of the regularization parameters α, λ is the fundamental part.
- Selecting well is essential to the performance, since it controls the strength of shrinkage and variable selection, which, in moderation can improve both prediction and interpretation.
- However, if the regularization becomes too strong, important variables may be left out of the model and coefficients may be shrunk excessively, which can harm both predictive capacity and the inferences drawn about the system being studied.

FORDHAM
THE JESUIT UNIVERSITY OF NEW YORK

Gabelli School of Business

Information Criteria, LASSO and Cross Validation

Information Criteria, LASSO and Cross Validation

- We set the parameter $\alpha = .5$, and estimate the coefficients of the model for alternative values of λ.
- As λ increases, more and more parameters go to zero.
- One way to choose this parameter is to use a method based on Cross Validation.
- We set the parameter $\alpha = .5$, and estimate the coefficients of the model for alternative values of λ.
- In this approach, we select a grid of values for λ, between $\lambda = 0$, and λ^*, the minimum λ which sets all of the coefficients $\beta_i = 0$.
- We then select a set of out-of-sample Mean Squared Error measures, based on holding out 20% of the sample for each specified λ over the grid.
- We thus select the optimal λ as the one which minimizes the average out-of-sample mean squared error, based on five sets of hold-outs of 20% of the data.
- We note that the coefficients $\{\beta_i\}$ are based on the full in-sample elastic-net estimation with the pre-specified tuning parameter, α,
- We obtain the final optimal value of λ from the cross-validation method.

Information Criteria, LASSO and Cross Validation

- We estimate he coefficients in the following steps
  1. specify $\alpha = .5$ for the elastic-net estimation, as a fixed hyper-parameter;
  2. full sample elastic-net estimation with various λ;
  3. cross validation with various λ;
  4. choose the optimal result based on the average mean-squared out-of-sample errors.
- We note that the coefficients $\{\beta_i\}$ are based on the full in-sample elastic-net estimation with the pre-specified tuning parameter, α, and the final optimal value of λ, coming from the cross-validation method.
- We also note that one can alter the Withholding Percentage from 20% to 10%. However 20% is widely used.
- We also note the Elastic Net with Cross Validation is ruthless: it kills off a lot of parameters.

FORDHAM
THE JESUIT UNIVERSITY OF NEW YORK

Gabelli School of Business

Example with Monte Carlo Simulation

Example with Monte Carlo Simulation

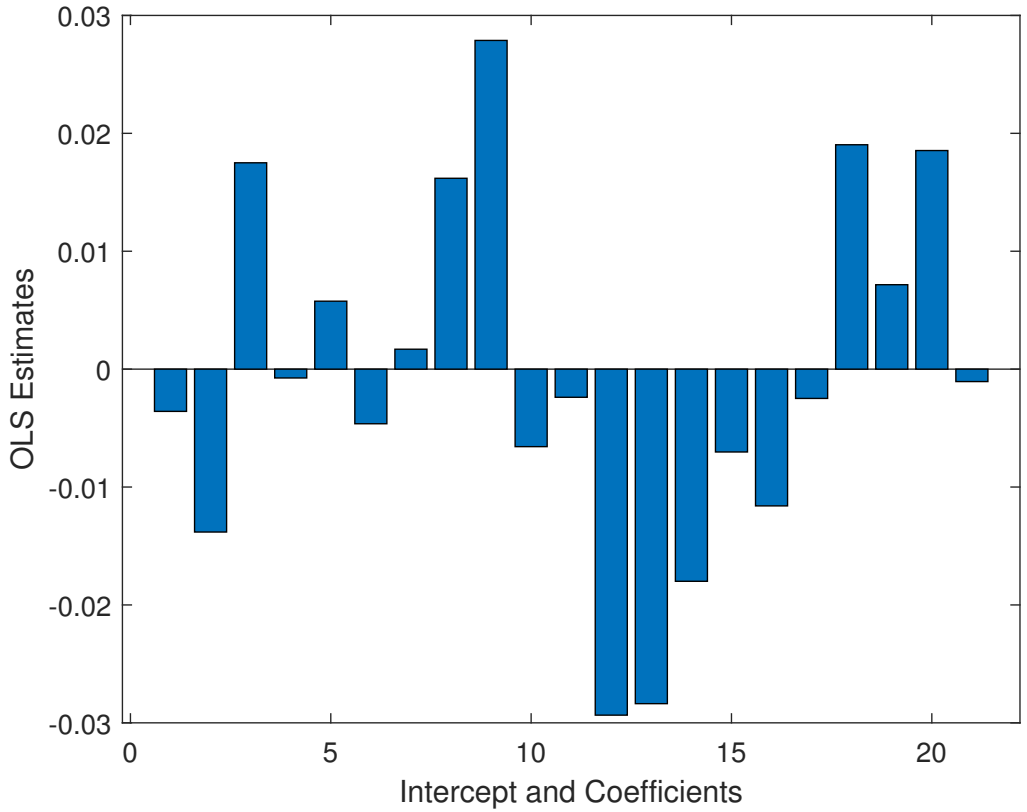
- In this example we generate random data series with a target variable y and twenty covariates x.
- They are independent and random, normally distributed with a mean zero and a standard deviation of unity.
- So if we do a regression we should get nothing

$$y_i = \beta_0 + \sum_{i=1}^{20} \beta_k x_{k,i} + \epsilon_i \tag{8}$$

- The constant term α and the coefficients β_k,$k = 1$,...20 should all be zero, for sure.
- When we do a regression of y and the constant term and the ten covariates, we get all sorts of numbers
- We call these nuisance parameters. Of course they are not statistically significant. But the OLS is trying to fit the data.
- And of course, random number generators are not totally random. How can write a computer program to generate randomness? The numbers are **psuedo-random**.

Example with Monte Carlo Simulation

- If we do a regression, illustrated in a Jupyter notebook, for this Monte Carlo experiment, we get OLS coefficients:



- If we are not interested in using T-statistics and want to use the model for prediction, we need to find a way to get rid of these coefficients, wipe them out. They represent "white noise". We should not pay attention to them.

- Lasso to the Rescue: for Elastic Net ($\alpha = 1$) in equation 7, with Cross Validation for 10 folds (each time withholding 500 observations and minimizing the out sample errors), with $\lambda = .0293$, all of the coefficients are driven to zero.

- If we set $\alpha = .5$,for equation 7, and with a combination of LASSO and Ridge Regression, the $\lambda = .0585$. All of the coefficients disappear.

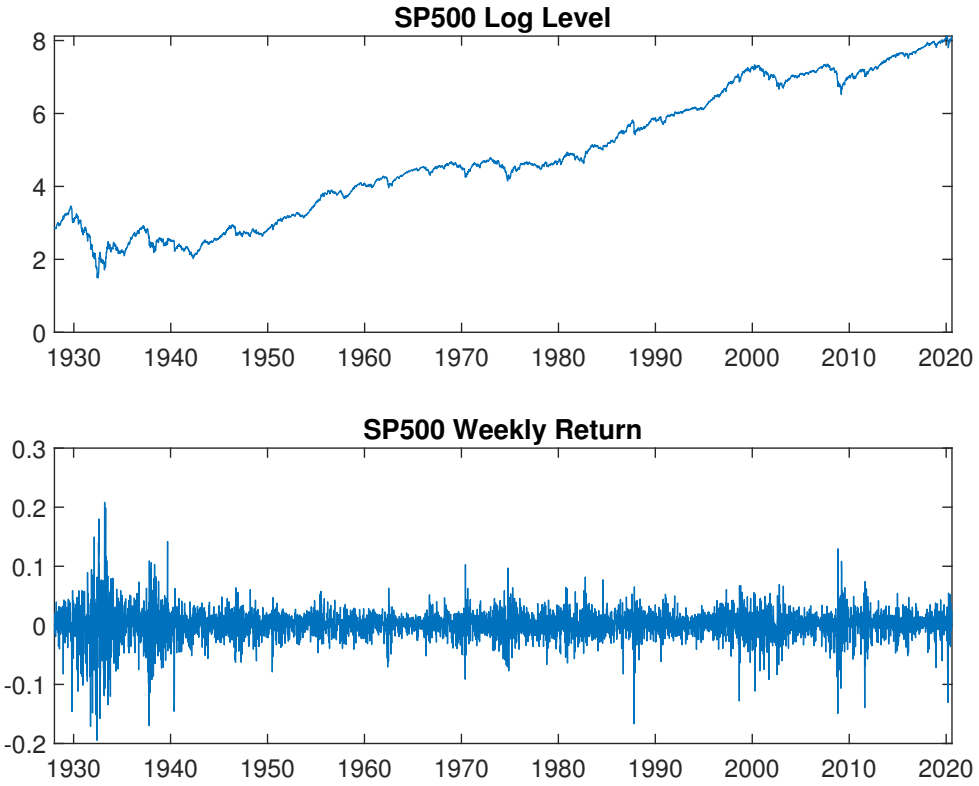
FORDHAM
THE JESUIT UNIVERSITY OF NEW YORK

Gabelli School of Business

Forecasting SP 500 Weekly Returns

Forecasting S&P 500 Weekly Returns

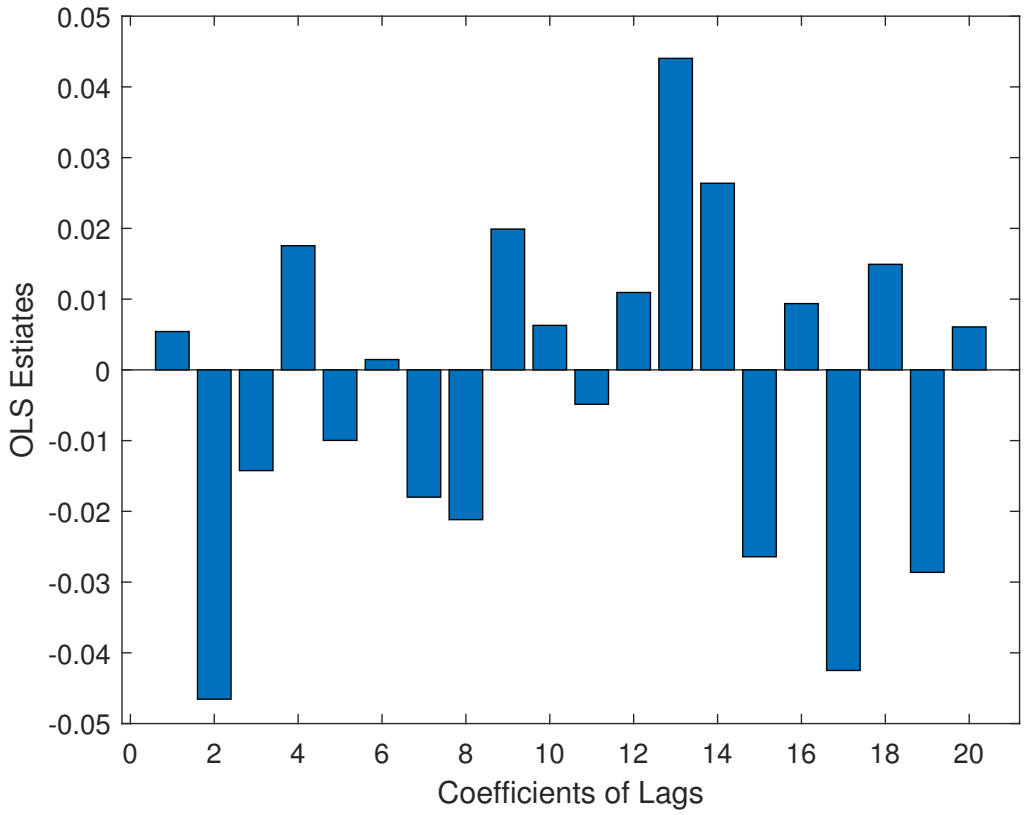
- Let's look at an empirical example: forecasting the weekly return on the S&P 500..



- You can see the steady upward climb and recurring periods of high volatility in this index.
- Lets compare the forecasting performance of least squares regression and LASSO.

Forecasting S&P 500 Weekly Returns

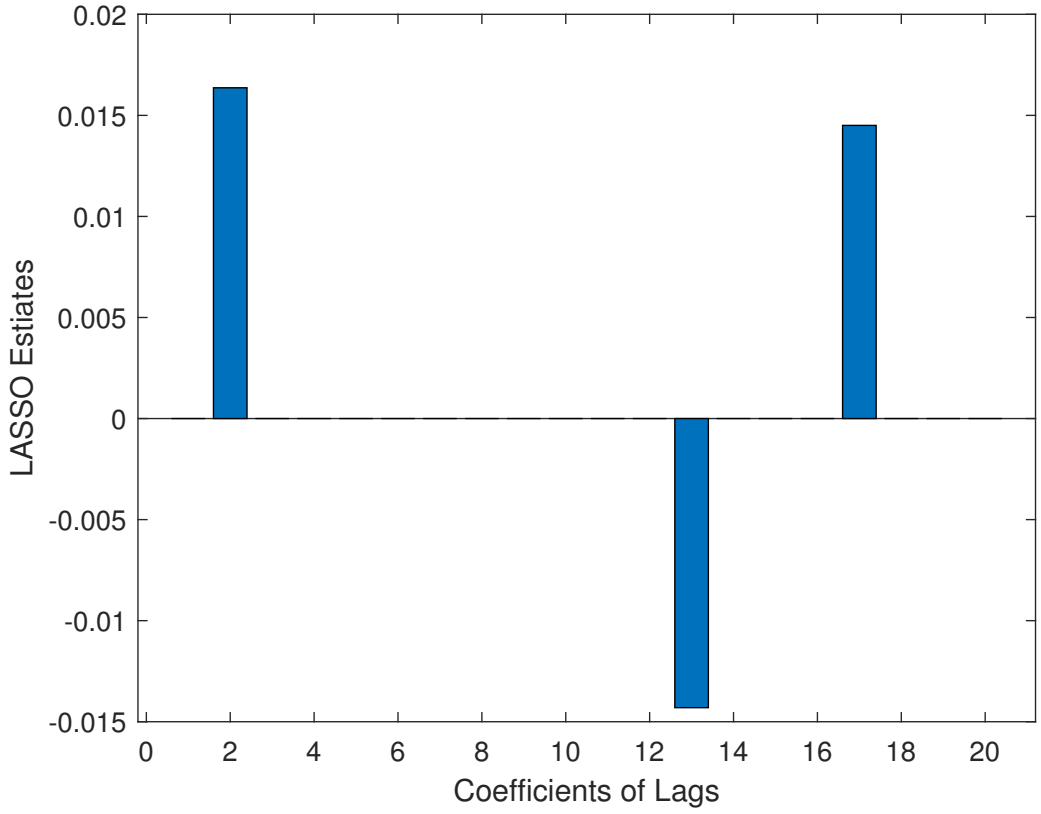
- We picture the estimated coefficients of an AR(20) model for the S&P 500 Weekly Returns.



- The key question: do we need to look back for 20 weeks in order to get a good forecast of the SP500 returns?

Forecasting SP 500 Weekly Returns

- LASSO to the rescue! $\alpha = .5, \lambda = .0015$,we get rid of most but not all of the coefficients.



- LASSO economizes on the information needed for prediction.

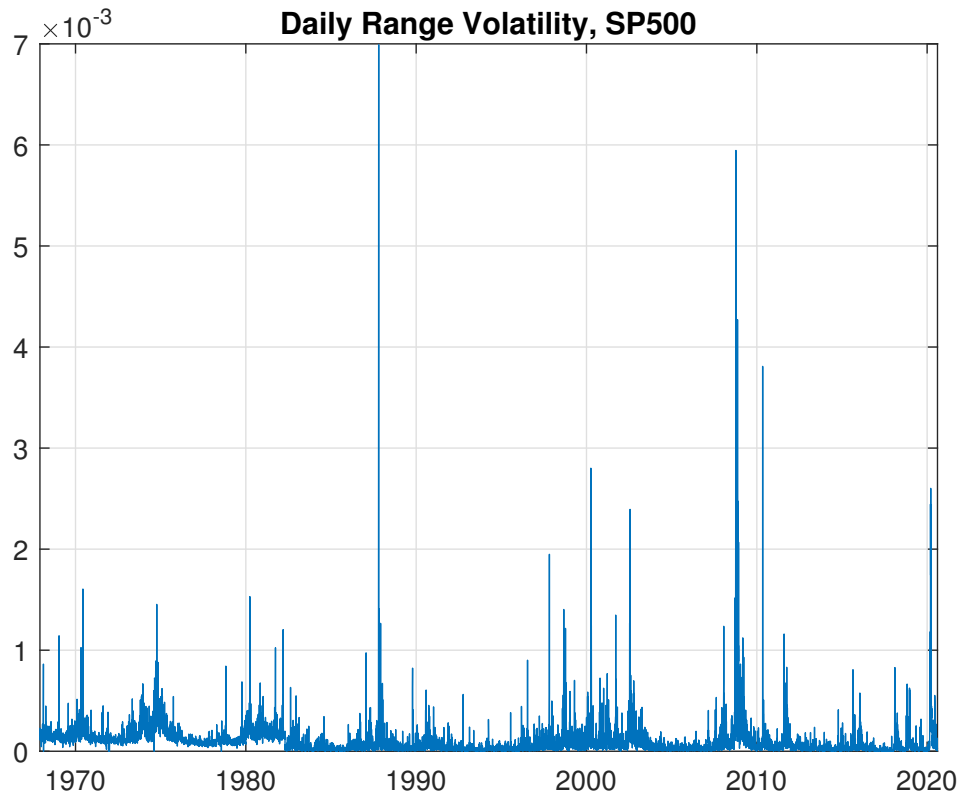
FORDHAM
THE JESUIT UNIVERSITY OF NEW YORK

Gabelli School of Business

Forecasting Daily Volatility of the SP 500 Index

Forecasting Daily Volatility of the S&P 500 Index

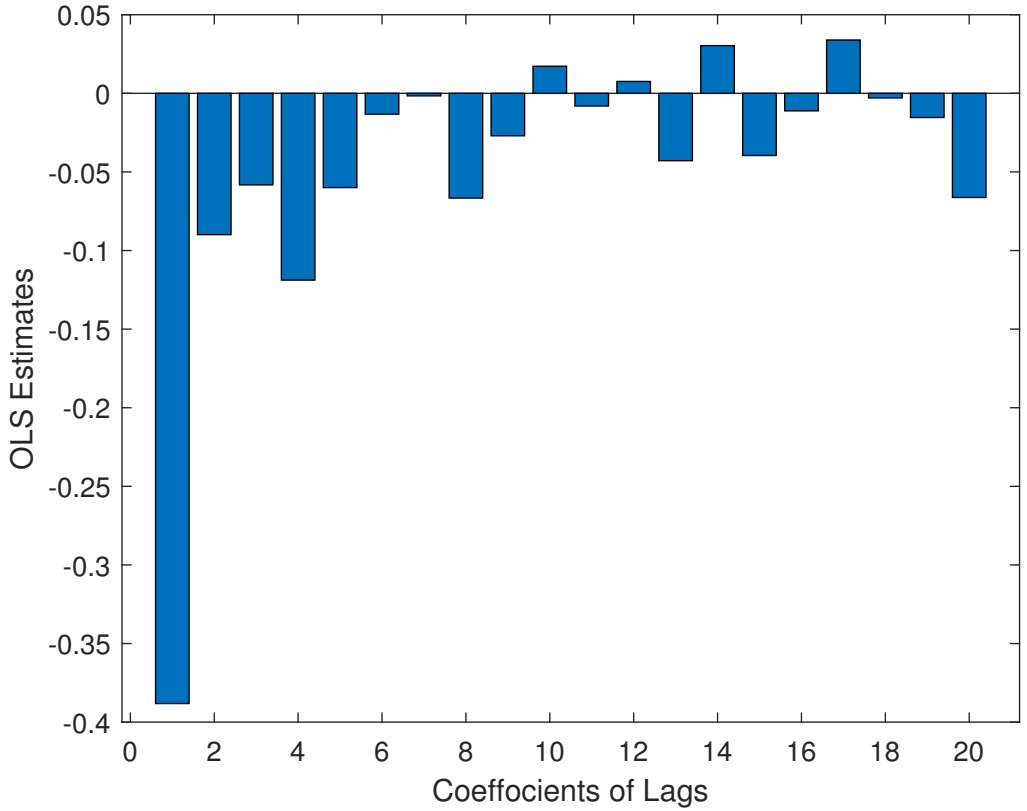
- Now lets examine the Range Volatility calculated from the daily open, high, low and closing estimates of the S&P 500 Index.
- Range Volatility Measure for the SP500 between Dec. 1967 and Aug. 2020:



- We see that there are spikes in volatility corresponding to key events,

Forecasting Daily Volatility of the S&P 500 Index

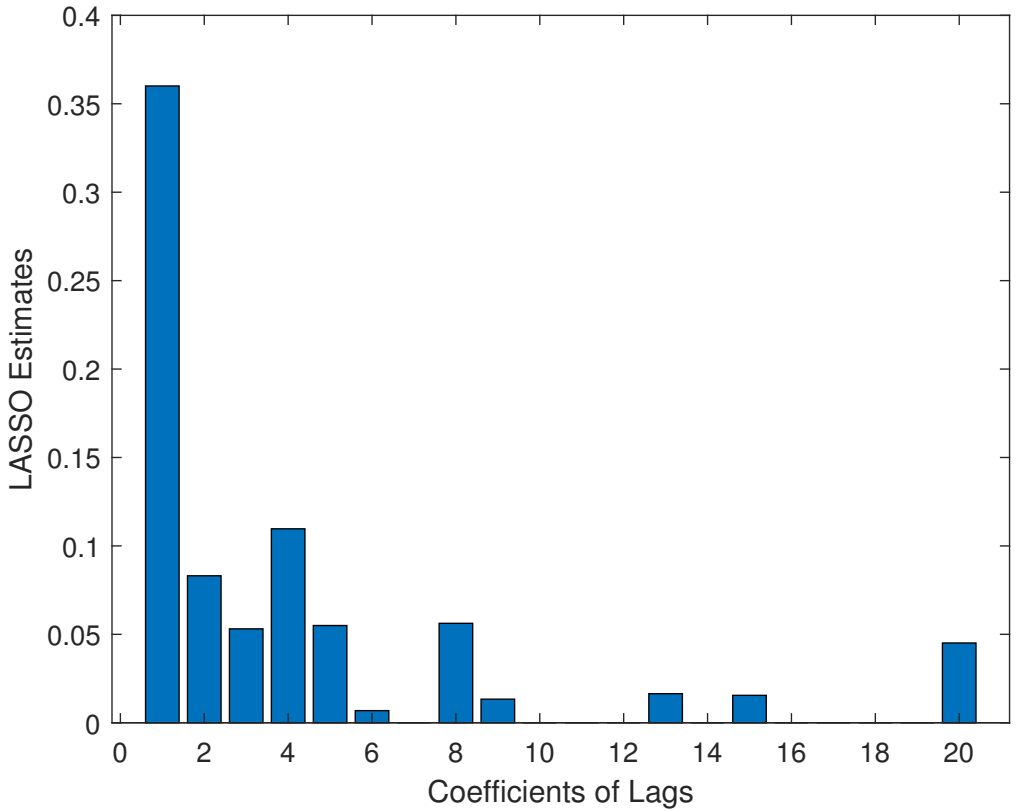
- OLS Estimates, with Lags of 20 (one month of business days)



- We see a considerable degree of negative serial correlation, what goes up comes down and what goes down comes up. This is a case of fitting the data with in-sample error minimization.

Forecasting Daily Volatility of the SP 500 Index

- LASSO tells another story. The process now shows positive serial correlations. Several longer lags are eliminated.



FORDHAM
THE JESUIT UNIVERSITY OF NEW YORK

Gabelli School of Business