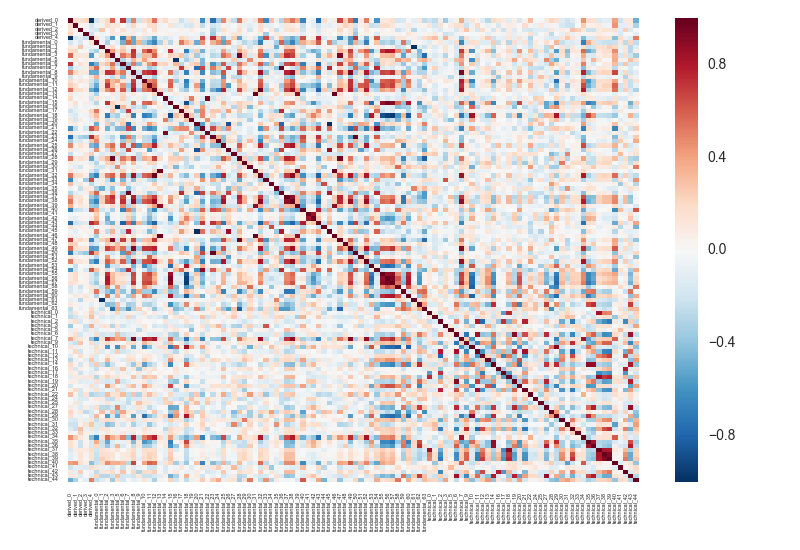
# Results of Principal Component and Ridge Regression

Pair-wise correlation among features over the entire training set:



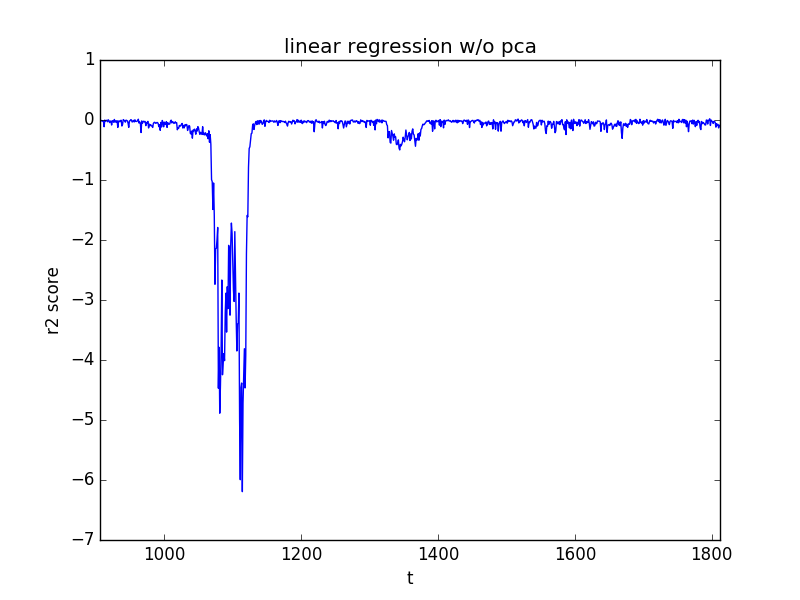
The abundance of collinearity among features implies pure linear regression may not work.

This problem is addressed by two methods: principal component and ridge regression.

To assess the performance of each, I divide the data sets into 60%:40% and use the first portion as training and the second as testing.

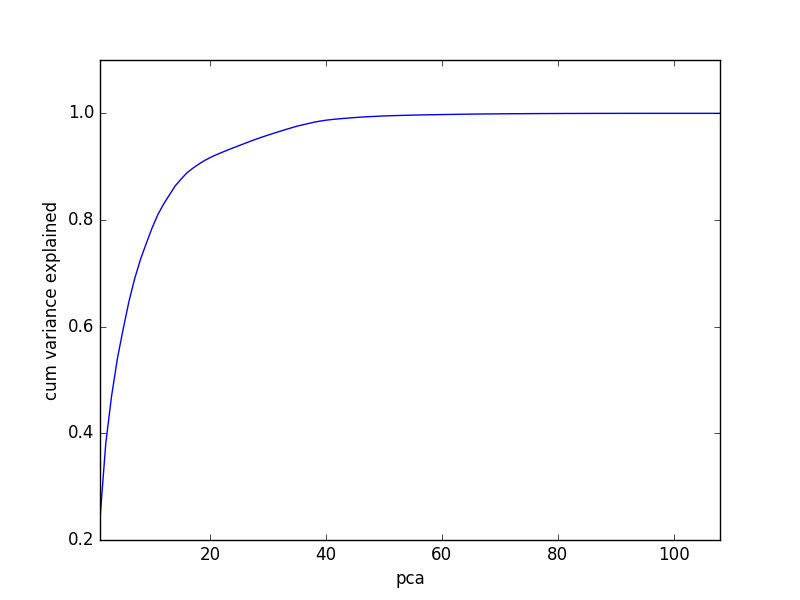
A time series of r2 scores is generated from the testing set: the best score is 1 and the worst score can be arbitrarily negative.

Below is the time series of r2 scores with pure linear regression as our baseline:



## Principal Component Analysis

Variance explained with Principal Component: 17 PCs explains ~ 90% variance



Time series of r2 scores of testing set using different # PCs: Overall scores have improved by reducing # PCs and stabilize at ~ 17 PCs.

|  |  |
| --- | --- |
|  |  |
| **All PCs** | **60 PCs** |
|  |  |
| **17 PCs** | **2 PCs** |

Statistics of r2 scores using 17 PCs:

|  |
| --- |
| count 907.000000 |
| mean -0.029043 |
| std 0.038069 |
| min -0.264366 |
| 25% -0.036402 |
| 50% -0.015020 |
| 75% -0.004523 |
| max 0.003937 |

## Ridge Regression

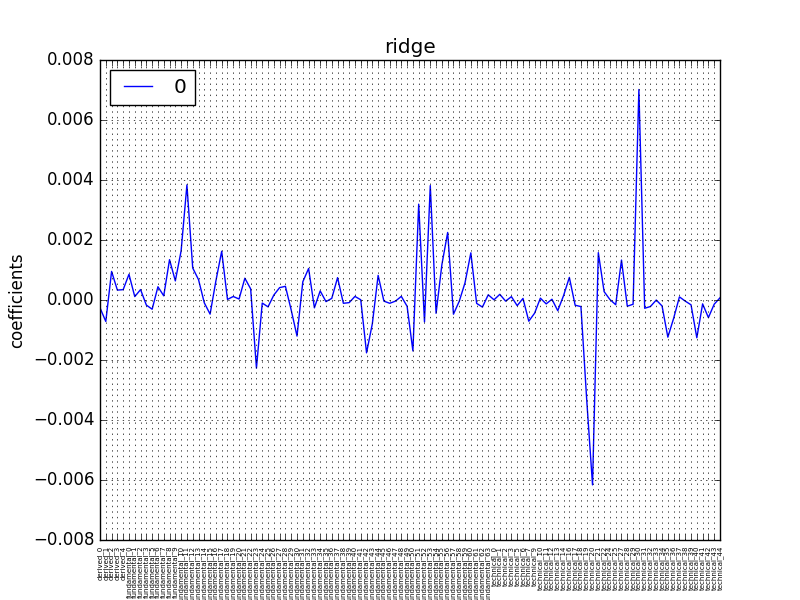
Time series of r2 scores of Ridge Regression using different alphas (penalty coefficient): scores have improved by increasing alpha. But alpha cannot be too big or else the model will become trivial.

|  |  |
| --- | --- |
|  |  |
| **Alpha = 1** | **Alpha = 50** |
|  |  |
| **Alpha = 100** | **Alpha = 200** |

Statistics of r2 scores with alpha = 200:

|  |
| --- |
| count 907.000000  mean -0.061561  std 0.117495  min -0.842012  25% -0.059893  50% -0.023498  75% -0.009553  max 0.016547 |

Coefficients of features with alpha = 200:



Features ranked by top coefficients are listed below: ‘technical\_20’ and ‘technical\_30’ stand out much more than the other features.

|  |  |
| --- | --- |
| **Feature Name** | **Coefficient** |
| ‘technical\_20’ | -0.006 |
| ‘technical\_19’ | -0.003 |
| ‘fundamental\_23’ | -0.002 |
| ‘fundamental\_42’ | -0.002 |
| ‘fundamental\_50’ | -0.002 |
| ‘fundamental\_56’ | 0.002 |
| ‘fundamental\_51’ | 0.003 |
| ‘fundamental\_53’ | 0.004 |
| ‘fundamental\_11’ | 0.004 |
| ‘technical\_30’ | 0.007 |

## Conclusions

PC and Ridge Regression have performed better than pure linear regression.

PC performs on average better than Ridge but the latter has a significantly better max score, i.e., 0.02 (ridge) vs. 0.004 (PC).

‘technical\_20’ and ‘technical\_30’ may be of special interest and for further analysis please refer to my Jupyter Notebook “Analysis of technical\_20 and technical\_30’.