Outlier Detection based on LOESS

LOESS: local regression/ locally weighted polynomial regression

**Basic idea:**

For each point in the series, fit a low-polynomial regression to the subset of its neighbor data using least squares, and give them more weights if nearer to the response point.

**Procedure and some finding:**

* The example is performed on INR266
* Since different rate has different length, I currently use the data up to 2010-03-22 to include all 13 curves.

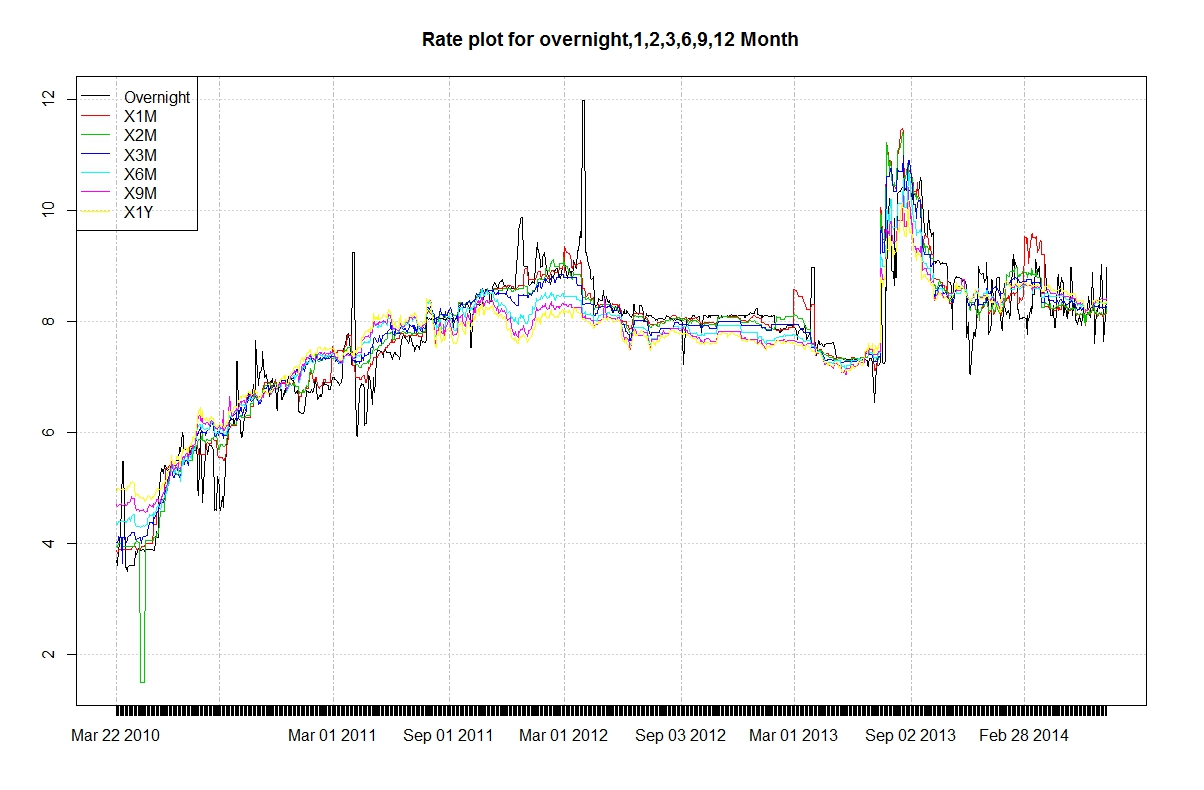
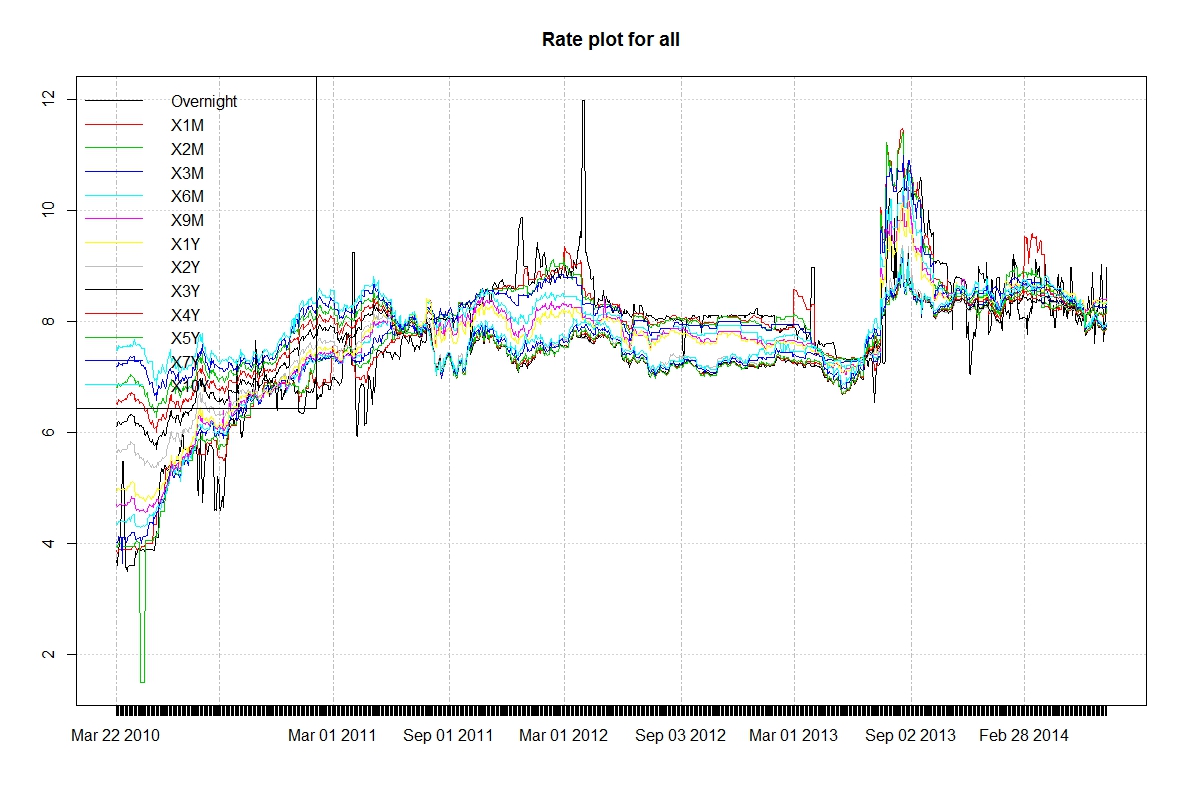
1. You may check the series plot for all 13 curves in appendix first. Obviously they are closely correlated and the volatility of each curve seems increase with the maturity. But the overnight rate seems to be most volatile. And one interesting feature is that the curves tend to converge from the past to the present.
2. For each curve, run LOESS with span = 0.3 (span is a parameter ranging from 0 to 1, lower span means higher focus on the local behavior). Plot the original series, fitted curve and residual curve (see appendix).
3. Classify the points with residuals out of range \big[ Q_1 - k (Q_3 - Q_1 ) , Q_3 + k (Q_3 - Q_1 ) \big] (Q\_1 and Q\_3 are the lower and upper quartiles respectively, k is the selected threshold) as the outliers. The outlier is marked in plots in appendix. Here I choose k as 1.96.

**Remarks:**

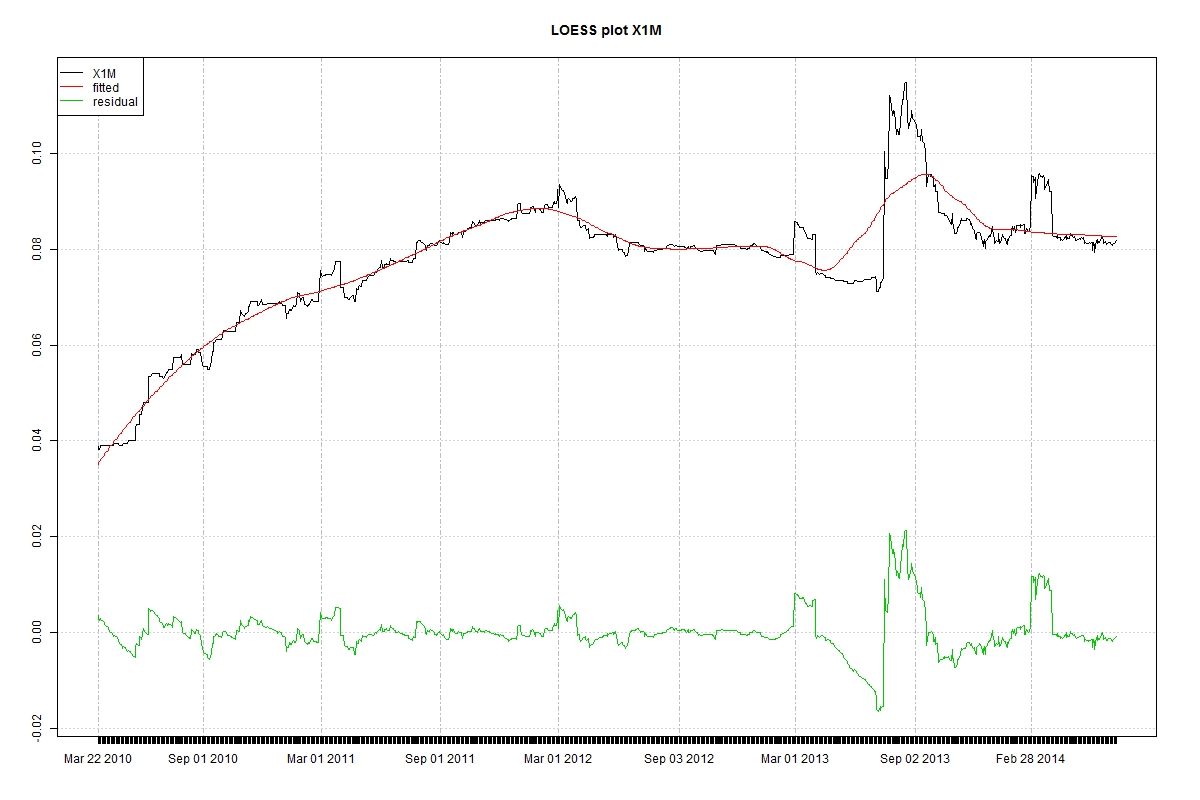
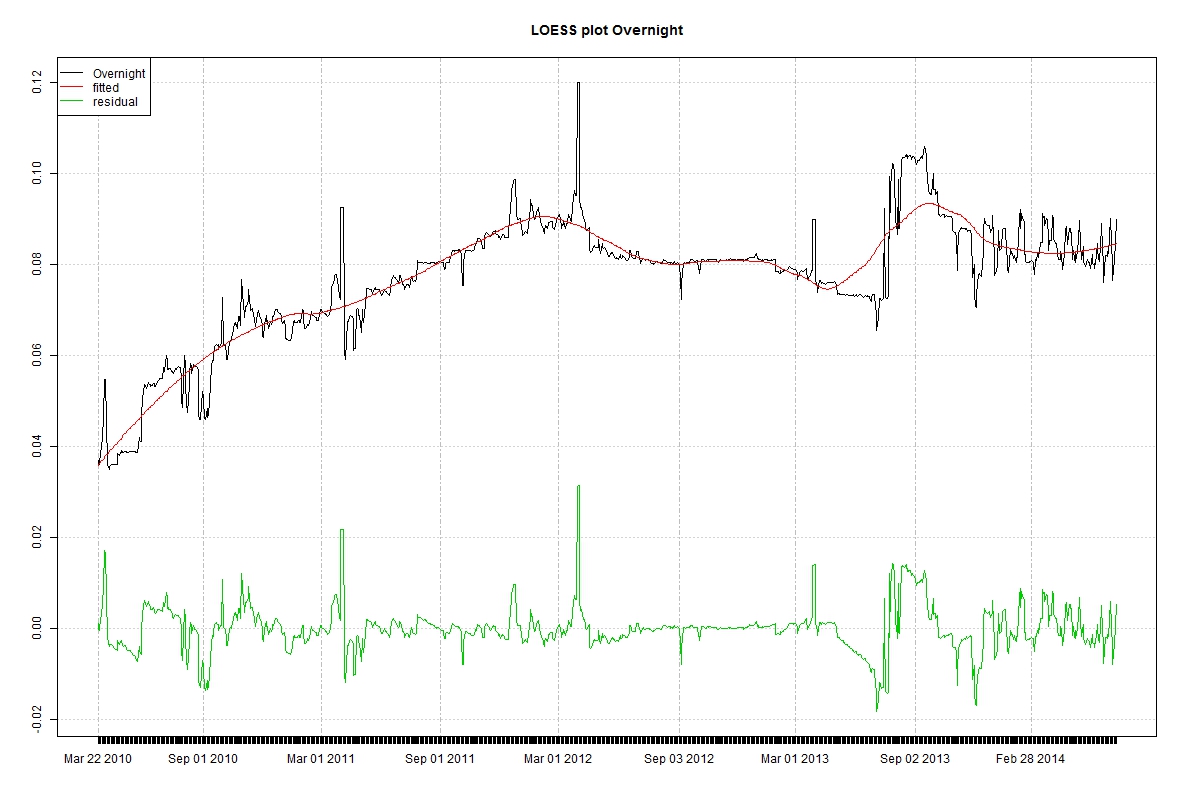
* LOESS does not specify a model to fit the whole series, it quite robust.
* Here I do not consider the collectively behavior of the near curve. We may cross check the outliers between each curve for the potential candidate date.
* For curve with more volatility, since many of its residuals are large (which means the absolute value of Q1 and Q3 tend to be larger), we will find very few outliers by using the interquartile method. Thus we may consider other criteria.
* Some detected possible outliers tend to be clustered. Are clustered outliers still outliers?
* I perform PCA (principal component analysis) on the 13 curves. The result shows that the first 2 components can explain over 98% of the whole variance within the data set. But not sure how to utilize this finding.

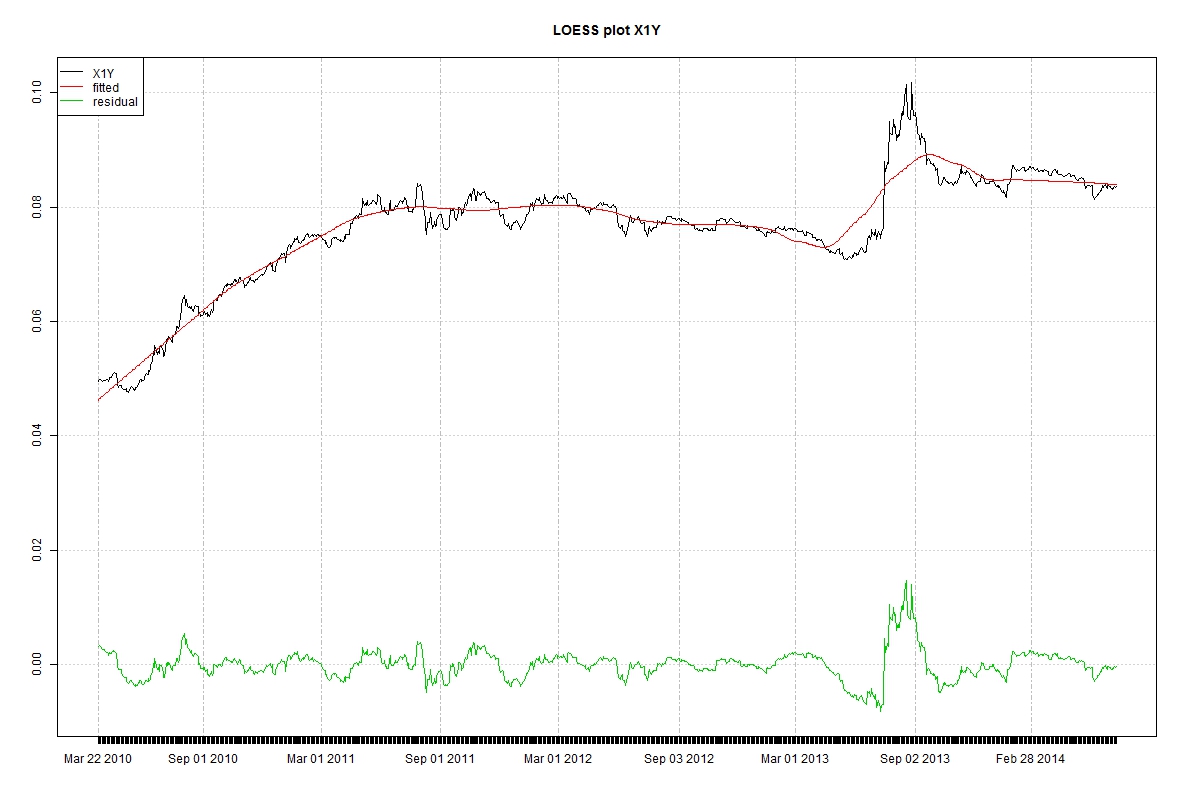
**Appendix:**

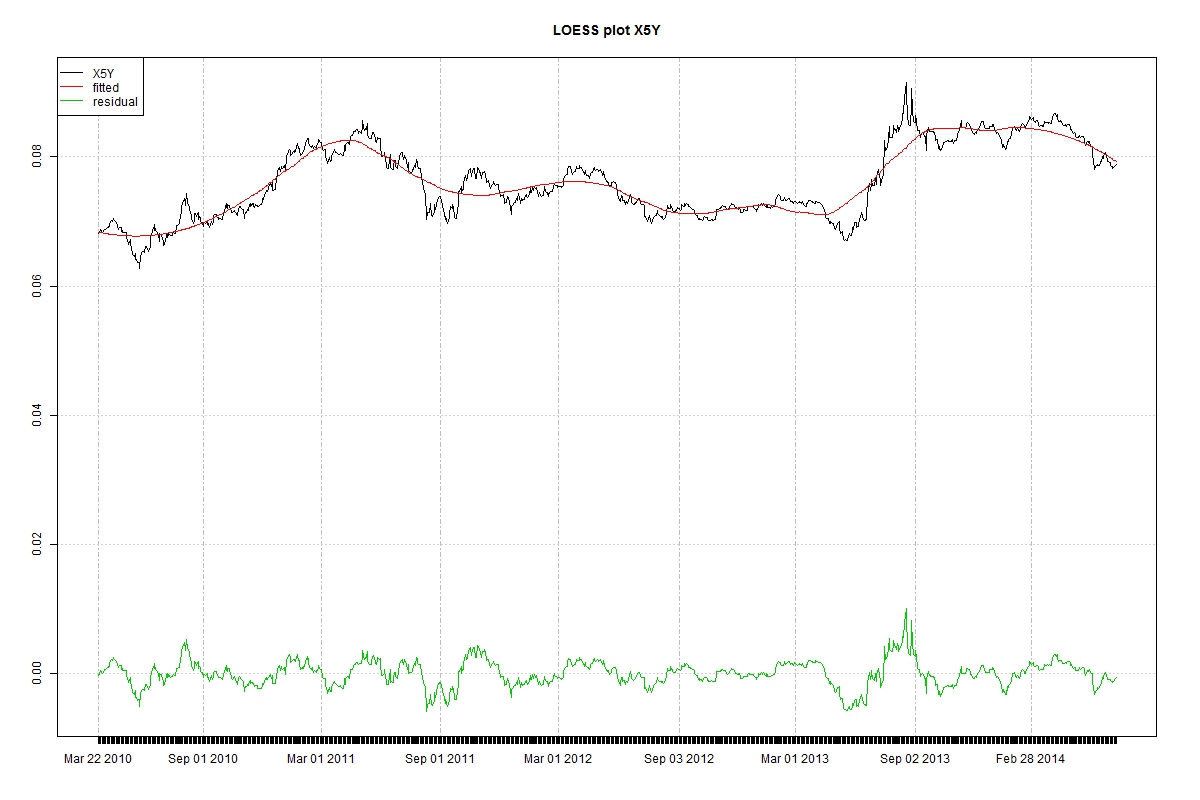
1. Original Rate plot

 2. LOESS fitting and residual (selected)

- the time horizon is labeled in the plot title.







Outlier plot (selected)

