Google Trends 波動度預測 研究計畫

曾貫岑

1. 標的 (2015/01/01-2020/07/07 日資料)台指期貨、A50指數期貨、星日經指數期貨、小SP指數期貨、DAX指數期貨、FTSE 100指數期貨、CAC40指數期貨。

|  |  |
| --- | --- |
| 標的期貨 | Google Trends關鍵字選取 |
| 台指期貨(TXF) | '台指期貨','期貨','台指期指數','道瓊指數','美股',  '台指期即時','台指期夜盤指數','台指結算','台指期貨夜盤',  '台指選擇權','台指期結算','台指保證金','台股指數',  '台指期夜盤即時','台指電子盤','摩台結算','摩台指結算' |
| A50指數(CA) | 'a50','a50指數','a50指數期貨','a50期貨','中國a50指數',  '富時中國a50指數','china a50 index',  'ftse china a50 index','a50 china',  '富时中国A50指数','ftse china' |
| 星日經指數(EN) | '日経平均株価','株価 日経 平均 株価','日经平均指数',  '日経 平均','Nikkei 225 Index','nikkei 225','nikkei index' |
| 小SP指數(ES) | 's&p 500','s&p 500 index','s&p500 stock',  's&p 500 futures','s&pfutures','s&p 500 price',  'the s&p 500','etf s&p 500','s&p 500 today',  's&p 500 chart','s&p 500 stock price' |
| DAX指數(DA) | 'wie steht der dax','wie steht der dax heute',  'wie steht der dax aktuell','dax aktie',  'dax heute','der dax heute','der aktionär',  'börse','dax aktuell','dax kurs','dax 30',  'dax realtime','dax','德国DAX指数' |
| FTSE 100指數(FT) | 'ftse 100','ftse 100 price','ftse 100 share price',  'ftse 100 index','ftse 100 news','ftse 250',  'bbc ftse 100','ftse 100 uk','ftse 100 bbc news',  'ftse 100 live','ftse 100 today','the ftse 100' |
| CAC40指數(CF) | 'CAC 40','cac 40 cours','cac 40 bourse','bourse',  'action cac 40','cac 40 boursorama','boursorama',  'cac 40 direct','le cac 40',  'cac 40 france ','cac 40 en direct' |

1. Google Trends 關鍵字設置

|  |  |
| --- | --- |
| keywords\_fix | 'debt', 'color', 'stocks', 'restaurant', 'portfolio', 'inflation', 'housing', 'dow jones', 'revenue', 'economics', 'credit', 'markets', 'return', 'unemployment', 'money', 'religion', 'cancer',  'growth', 'investment', 'hedge', 'marriage', 'bonds',  'derivatives', 'headlines', 'profit', 'society', 'leverage',  'loss', 'cash', 'office', 'fine', 'stock market', 'banking',  'crisis', 'happy', 'car', 'nasdaq', 'gains', 'finance', 'sell', 'invest', 'fed', 'house', 'metals', 'travel', 'returns', 'gain',  'default', 'present', 'holiday', 'water', 'rich', 'risk',  'gold', 'success', 'oil', 'war', 'economy', 'chance',  'short sell', 'lifestyle', 'greed', 'food', 'financial markets', 'movie', 'nyse', 'ore','opportunity', 'health','short selling', 'earnings', 'arts', 'culture', 'bubble',  'buy', 'trader', 'rare earths', 'tourism', 'politics',  'energy', 'consume', 'consumption', 'freedom',  'dividend', 'world', 'conflict', 'kitchen', 'forex',  'home', trash', 'transaction', 'garden', 'fond', 'train',  'labor', 'fun', 'environment', 'ring' |
| Domestic trend | 'advert','airtvl','autoby','autofi','bizind','bnkrpt','comput','crcard','durble','educat','invest'  ,'finpln','furntr','insur','jobs','luxury','mobile','mrtge','rlest','rental','shop','smallbiz','travel' |

1. **研究方法**  
   (1) 爬蟲抓取關鍵字(2015/01/01-2020/07/07 日資料)  
    每次最多能爬18萬筆資料。  
   (2) 資料前處理  
    a.若關鍵字熱度指數超過10% 遺漏值，則刪除此關鍵字。  
    b.關鍵字熱度指數過於極端(大部分為0，特定時間段為100)，則刪除  
    此關鍵字。  
    c.設置波動度:True Range、日內波動度sigma(German ,1980)  
    d.設置inputs: Open、High、Low、Close、Volume、return、sigma  
    、TR、關鍵字熱度(標的名稱加上crossmean，約110-120個)。

(3) 資料分析

觀察各變數與波動度的散佈圖，比較F-statistics、Mutual Information。

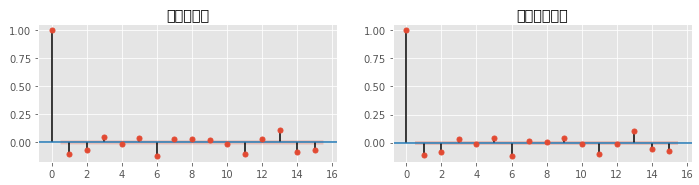
觀察極端值or波動度小的分布。

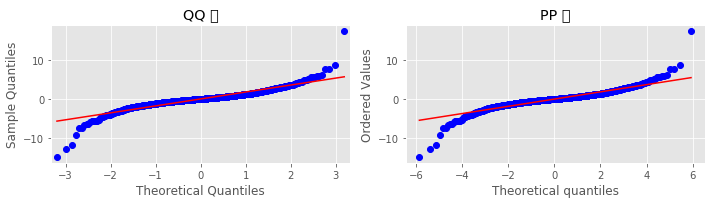
(4) 配飾模型:  
 A. GARCH、EGARCH、TGARCH…  
 B. LSTM、GRU…  
(5) 比較結果

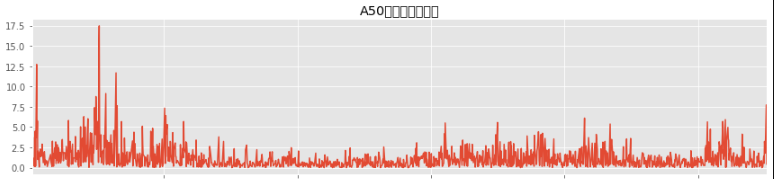
1. **研究結果**
2. **A50**配飾GARCH類模型

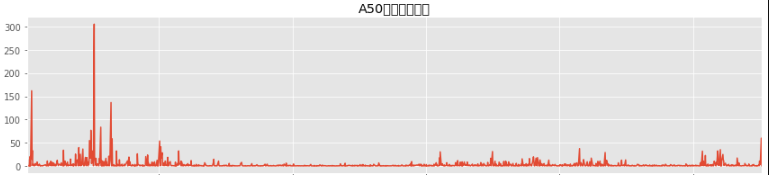
觀察右圖A50歷史日報酬率

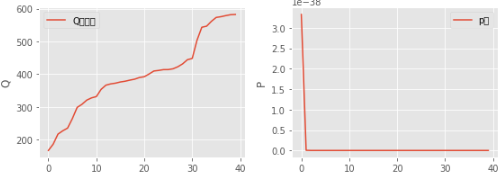
圖二、三為報酬率的ACF、PACF圖，自我相關比較顯著的有前一、二、六期。



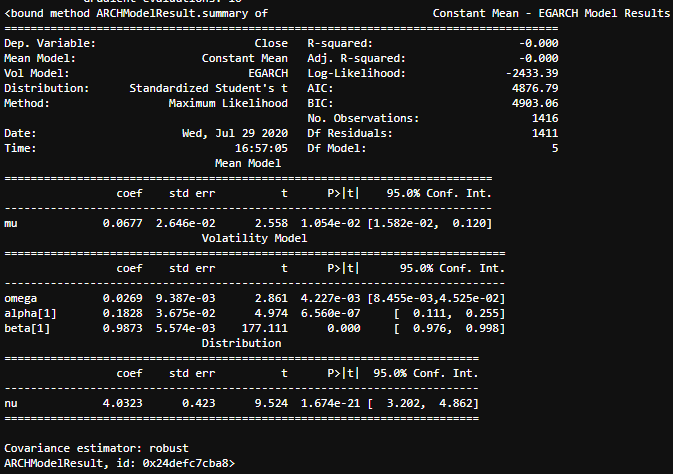
觀察QQ圖和PP圖，應為厚尾分布。

觀察報酬平方項(上圖)和絕對值項(下圖)，有波動群聚的現象。

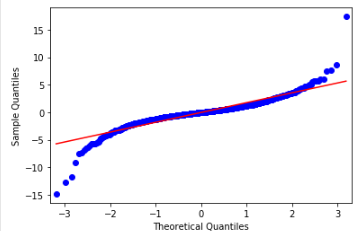


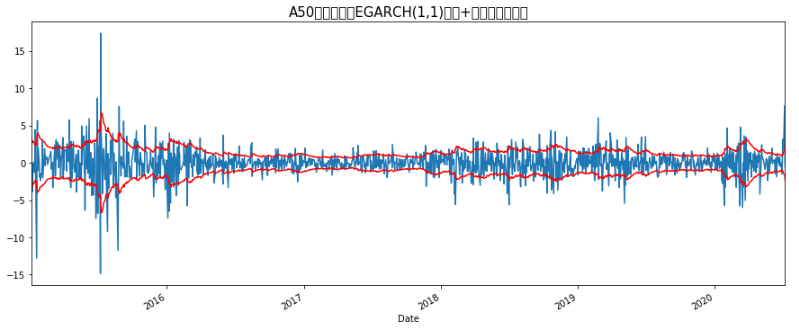
做ljun-box檢定，以p-value看非常顯著。

嘗試不同GARCH模型，以AIC、BIC為標準，**EGARCH + student-t分配** 配飾的較好，且以係數的p-value來看皆為顯著，有顯著**波動度不對稱**的現象。



再觀察QQ圖，改用student – t分配後，還是呈後尾分布。



下圖為配飾模型狀況，配飾情況不佳，改用LSTM嘗試。

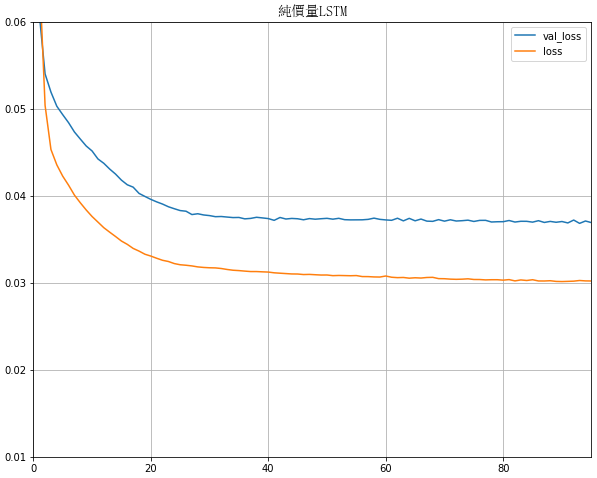
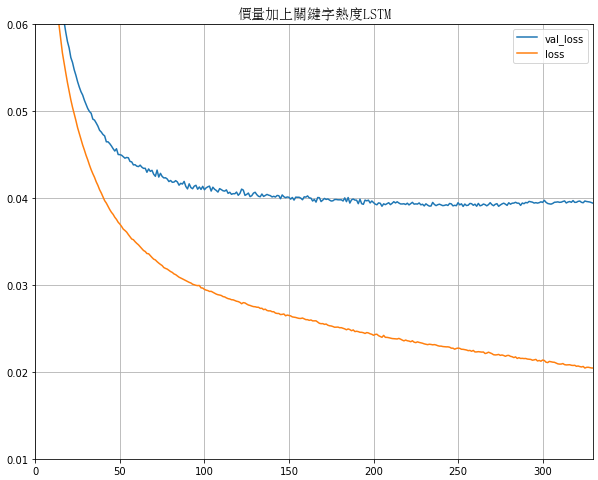
1. **A50配飾LSTM**
2. 設置input: Open、High、Low、Close、Volume、sigma、TR、return、標的名稱熱度的cross mean。
3. 分為預測sigma、TR。
4. 標準化
5. 隨機亂數，再分train、validation test data(90%、10%)。
6. 嘗試不同的Lag期數inputs(past days)、batch size、epochs。
7. 以validation MAE為衡量標準。

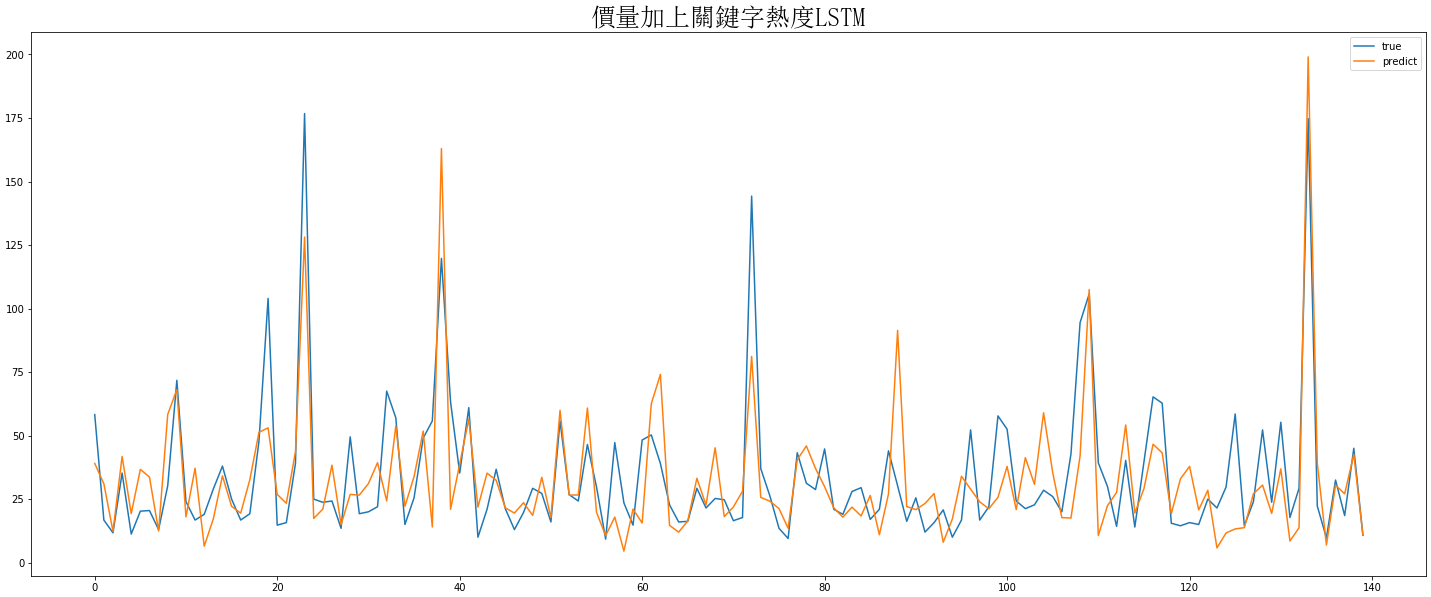
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **sigma** | **batch\_size** | **16** | **32** | **64** | **128** |
| **pastdays = 5** | epochs | val\_MAE | val\_MAE | val\_MAE | val\_MAE |
|  | 30 | 0.031114 | 0.031423 | 0.040502 | 0.047018 |
|  | 50 | 0.031222 | 0.030866 | 0.03102 | 0.038043 |
|  | 100 | 0.031314 | 0.031233 | 0.031131 | 0.031484 |
|  | 200 | 0.031133 | 0.030752 | 0.030905 | 0.031022 |
| **pastdays=10** | epochs | val\_MAE | val\_MAE | val\_MAE | val\_MAE |
|  | 30 | 0.03174 | 0.032958 | 0.032187 | 0.031529 |
|  | 50 | 0.030441 | 0.030721 | 0.030892 | 0.037745 |
|  | 100 | 0.031531 | 0.030482 | 0.03089 | 0.031036 |
|  | 200 | 0.030371 | 0.03094 | **0.030016** | 0.030929 |
| **pastdays=20** | epochs | val\_MAE | val\_MAE | val\_MAE | val\_MAE |
|  | 30 | 0.038137 | 0.040105 | 0.043296 | 0.040876 |
|  | 50 | 0.03768 | 0.039075 | 0.039968 | 0.039649 |
|  | 100 | 0.037106 | 0.03828 | 0.037586 | 0.038764 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **TR** | batch\_size | 16 | 32 | 64 | 128 |
| **pastdays = 5** | epochs | val\_MAE | val\_MAE | val\_MAE | val\_MAE |
|  | 30 | 0.03296 | 0.032631 | 0.04083 | 0.046986 |
|  | 50 | 0.032529 | 0.033247 | 0.03331 | 0.038489 |
|  | 100 | 0.032396 | 0.033064 | 0.032534 | 0.033612 |
|  | 200 | 0.032411 | 0.032516 | 0.032528 | 0.033196 |
| **pastdays=10** | epochs | val\_MAE | val\_MAE | val\_MAE | val\_MAE |
|  | 30 | 0.031331 | 0.031413 | 0.031791 | 0.046479 |
|  | 50 | 0.031207 | 0.031583 | **0.030481** | 0.034438 |
|  | 100 | 0.031157 | 0.031586 | 0.030789 | 0.031549 |
|  | 200 | 0.031447 | 0.031681 | 0.030951 | 0.030922 |
| **pastdays=20** | epochs | val\_MAE | val\_MAE | val\_MAE | val\_MAE |
|  | 30 | 0.040692 | 0.040341 | 0.041831 | 0.042101 |
|  | 50 | 0.040759 | 0.039295 | 0.039477 | 0.041852 |
|  | 100 | 0.039261 | 0.040352 | 0.040614 | 0.042867 |
|  | 200 | 0.03956 | 0.04018 | 0.0407 | 0.040125 |

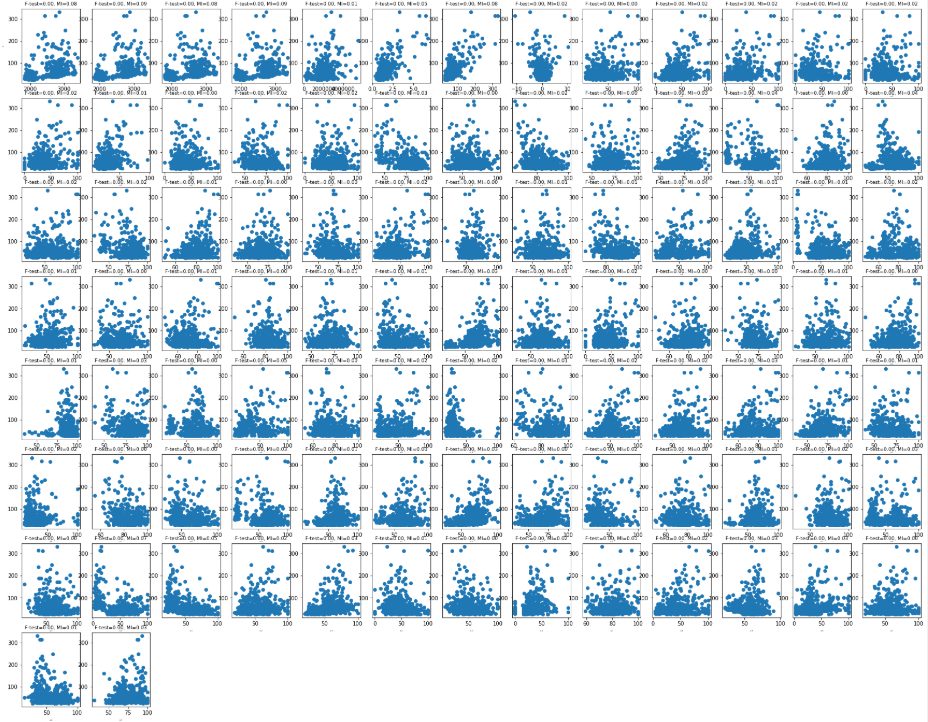
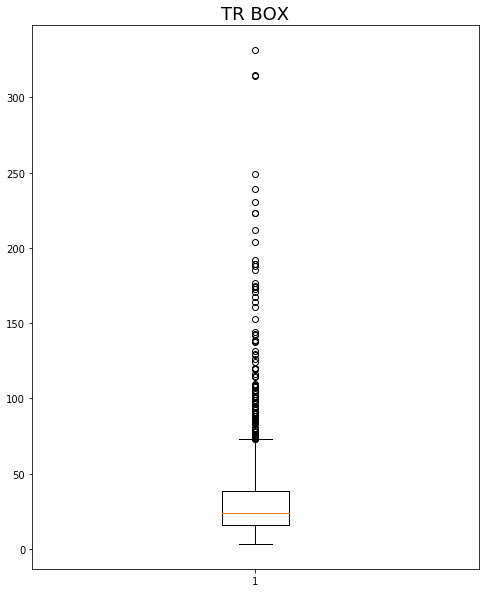
結果:

1. Batch size 設為64較合適，input的lag期數設為10(約兩個禮拜)為佳。
2. 將關鍵字熱度的變數拿掉後，發現validation loss並沒有太大差異，推論在A50指數Google Trends變數並沒有太多資訊。
3. 考慮Google Trends的地域關係，將標的改為S&P 500測試。
4. **小SP指數期貨配飾LSTM**
5. 設置input: Open、High、Low、Close、Volume、sigma、TR、return、標的名稱熱度 + 其cross mean、**Domestic Trends**、**keywords\_fix**。
6. 分為預測sigma、TR。
7. 觀察不同變數和波動度的散佈圖、Mutual Information。
8. 標準化。
9. 隨機亂數，再分train、validation test data(90%、10%)。
10. 嘗試不同的Lag期數inputs(past days)、batch size、epochs。
11. 以validation MAE為衡量標準。
12. 觀察波動度的分布，並嘗試拿掉波動大或波動小的部分，重複(a)-(g)步驟。





結果:

1. 比較有無Google Trends 關鍵字熱度變數的結果，發現加上熱度變數，validation loss反而上升，反而overfitting。
2. 以配飾結果來看，預測結果常低估實際情況。
3. 分別去看每個變數與TR的Mutual Information(下圖)，標準化後關鍵字熱度變數都非常低，作為模型的變數可能不合適。
4. 分別將波動大(前2.5%)、波動小(後50%)拿掉測試，  
   發現Google Trends熱度對波動大(前50%)的Mutual Information較高，但資訊量還是非常低。
5. 分別將波動大(前2.5%)、波動小(後50%)拿掉再配飾，  
   validation loss提高非常多。
6. 分不同時間段來觀察Mutual Information，都有所差異

，但資訊量都很低。

1. **結論與建議**