Fintech: From Research to Deployment

Philip L.H. Yu
Department of Statistics and Actuarial Science
The University of Hong Kong

HKU Fintech Symposium 2019

Jan 4, 2019

Fintech - Financial Technology

A new financial industry that applies technology to improve financial activities.

Source: Schueffel, P. (2017). Taming the Beast: A Scientific Definition of Fintech. Journal of Innovation Management. 4 (4): 32–54.

The FinTech revolution, driven by a wave of start-ups with innovative new business and revenue models, new products and services, is changing finance for the better globally.

Source: Chishti, S. & Barberis, J. (2016). The FINTECH Book: The Financial Technology Handbook for Investors, Entrepreneurs and Visionaries. Wiley.

InvestHK

Hong Kong FinTech 2018

Download and share

Expanding in Asia and looking for a base? Looking to tap the strengths of the Hong Kong FinTech ecosystem? At InvestHK we provide free and confidential business set-up facilitation. We're ready to help.
Contact us now at fintech@investhk.gov.hk

6 fasting-growing Fintech areas

- Artificial intelligence (AI)
- Blockchain
- Regtech
- Wealthtech
- Insurtech
- Cybersecurity

Automated Trading

Source: https://www.daytrading.com/

得大數據者得天下

數裏見真章
王緋、楊良河

近年來，大數據（Big Data）一詞時常出現在網路、報章雜誌甚至電視特輯中，甚是火熱。例如北美精算師協會在去年底出版的雙月刊就以「大數據：誰收集消費者數據？為什麼？」為封面主題。財爺於今年初宣讀財政預算案中，亦提及「資訊科技的迅速發展，將世界帶到指尖。」

■量化交易員利用大數據進行分析，極短時間內發掘金融市場價格波動的一些規律，從而及時調整自己的交易策略，並從中賺取利潤。（路透圖片）

處理和分析資訊的能力，成為現代大型企業競爭優勢的重要一環。……政府會研究進一步使用物聯網（Internet of Things）、感應器（sensors）和大數據分析（big data analytics）技術，更有效地管理我們的城市。」

邁入大數據時代

其實早在2012年初，美國政府已宣布一項「大數據研究與開發倡議」（Big Data Research and Development Initiative）計劃。包括美國科學基金會、能源部、國防部等六大部門宣布，投資超過2億美元，進行大數據研究以迎接大數據時代的到來和挑戰。這是繼20多年前美國政府宣布，「國家訊息基礎設施」（National Information Infrastructure, NII）計劃以來的又一個重大戰略計劃。可以說走過了20多年前興起的訊息公路時代，現代社會正邁入大數據時代。

大數據特色四個「V」

現今，隨着電腦的運算速度及儲存能力不斷提高以來，數據的儲存量一直迅速地增加。由以往只針對個別研究而收集數據，到現時在沒有特定的原因下，很多數據如客戶的購物地點和時間亦會一一儲存下來。及至近年互聯網與社交網絡服務的盛行如電子郵件、Google搜尋、Facebook留言、微博及WhatsApp短訊、Flickr照片上傳等，更令大數據以驚人的速度不斷膨脹。這些收集下來的大數據，亦反映出大數據四個「V」的基本特徵，即大量化（Volume）、快速化（Velocity）、多樣化（Variety）和真實性（Veracity）。

首先，大數據裏其中兩個的「V」是指數據量非常龐大（Volume）及數據產生的速度非常快（Velocity），龐大的數據使一般的電腦和傳統的處理方法，無法在合理的時間內對這些數據進行處理和分析。以熱門相片分享應用程式Instagram為例，平均每天使用者分享6000萬張新相片，即每分鐘要快速處理四萬多張相片。假設每張相片的大小平均有2MB，則每天instagram新增約105 TB（1TB=1024GB，1GB=1024MB）的相片。要知道現在市場一般個人電腦的硬碟容量只有平均500GB左右，讀者可以想像要處理這麼龐大的數據，可不是幾台電腦幾個小時就能處理完成的。

大數據的第三個「V」是指數據類型多種多樣及來源廣泛（Variety），除了傳統的文本的數據，大數據還包括圖片、音訊、視頻、地理位置、社交關係等等。大數據的來源更是非常廣泛，網路中人們的每一次點擊，金融業中每一毫秒的交易，氣象研究中每時每刻收發到的觀測訊息，以及生物領域數以億計的基因序列，都構成了數據的大量累積。

掌握投資者情緒得風氣之先

最後，大數據的真實性（Veracity）是指數據的可靠程度。例如一些網上評論不一定是全部真確，或是數據因某種原因而出現錯誤或異常。

根據一項由一家全球訊息服務集團Experian的調查結果，受訪機構認為高達17%的數據可能不準確。而27%的受訪機構不知道它們的客戶數據有多少是不準確的。

有人話「大數據是石油」，筆者覺得用石油來比喻大數據非常貼切。石油是很有價值的能源，石油能提煉出許多有用的化工產品如化肥、殺蟲劑和塑料等。同樣地，大數據對各行各業亦很有價值。例如金融業，大數據愈來愈頻繁地被用在高頻交易（High-frequency trading）和演算法交易（Algorithmic trading）中，以發現更多更有價值的交易機會。利用大數據進行分析的量化交易員（Quant Trader），在極短的時間內快速發掘金融市場價格波動的一些規律，從而及時調整自己的交易策略（Trading strategy），並從中賺取利潤。除了金融市場自身產生的大數據之外，交易員也可以同時分析上市公司發放的消息和環球經濟新聞，社交網路上人們對市場的討論及情緒變化等，歸納出一些對交易非常有指導意義的資訊，從而對市場走向作出更加準確的預測。

個人化推薦緊貼客戶需求

除了在金融業的應用，主要應用大數據技術的領域其實是電子商務，特別是客戶分析（customer analytics）上。此方面的佼佼者當然是亞馬遜（Amazon）。當一個人在亞馬遜網站瀏覽購物時，亞馬遜總是會給他推薦一些別的商品，而且這些推薦往往更能符合用戶的興趣，這就是個人化推薦（Personalized Recommendation）。所有使用者過往的龐大的網路記錄，包括瀏覽過的網頁、購買過的商品、對商品的評價及打分等，都是推薦常用的重要依據。除此之外，用戶之間的社交關係，商品之間的聯繫等，也能為用戶對商品的潛在興趣提供更準確的預測。

這些資料通常都是非常龐大並且相當複雜，如果能有效地分析這些大數據，提取出真正能反映使用者的購買意欲和興趣，則能有效地留住舊用戶並吸引更多的新的用戶。

最近，阿里巴巴舉辦了一個大數據競賽，比賽任務就是個人化推薦。利用用戶在阿里巴巴另一個購物網站「天貓」（Tmall）的行為日誌（包括點擊、購買、加入購物車、收藏4種行為），透過分析這些資料及建立模型去了解用戶的品牌偏好，並預測他們在未來一個月內對品牌下商品的購買行為。

由從事業務來看，阿里巴巴是一間電子商務的公司。但如果從交易的實質來看，阿里巴巴實際上是提供數據服務，利用大數據撮合交易賺錢！以上的大數據競賽，正好為阿里巴巴提供一些更準確預測購買行為的建模方法。競賽現仍在進行中，結果如何，讓者拭目以待。

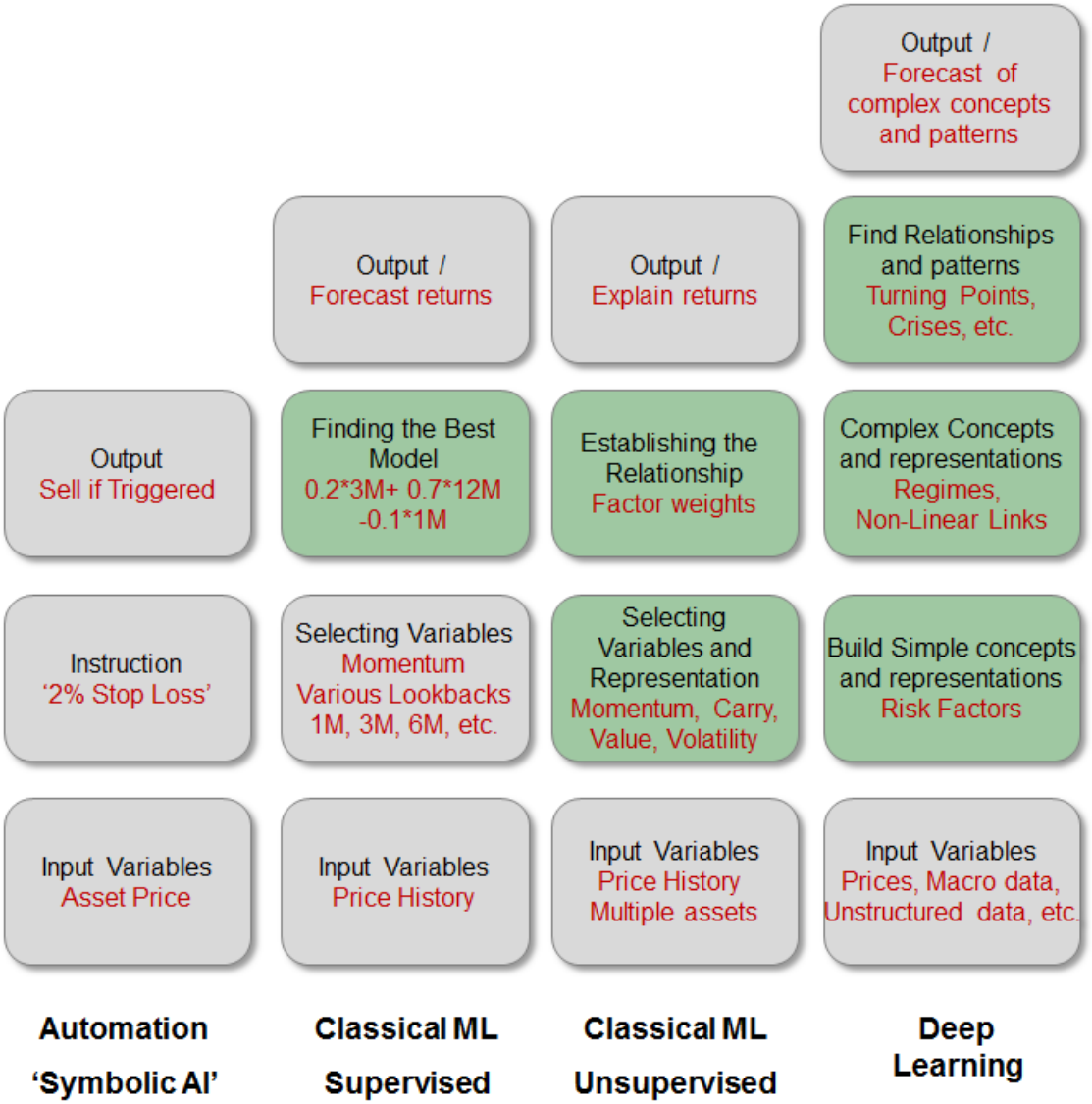
在大數據時代中，掌握大數據就是掌握機遇，關鍵是能否從大數據中挖掘出潛在的有效訊息。

毫無疑問，雲端運算（Cloud Computing）及統計數據挖掘（Data Mining）技巧是處理和分析大數據一個不可或缺的工具。篇幅所限，下期再續。

王緋為香港大學計算機科學系研究碩士畢業生，現任摩根士丹利投資銀行實習生

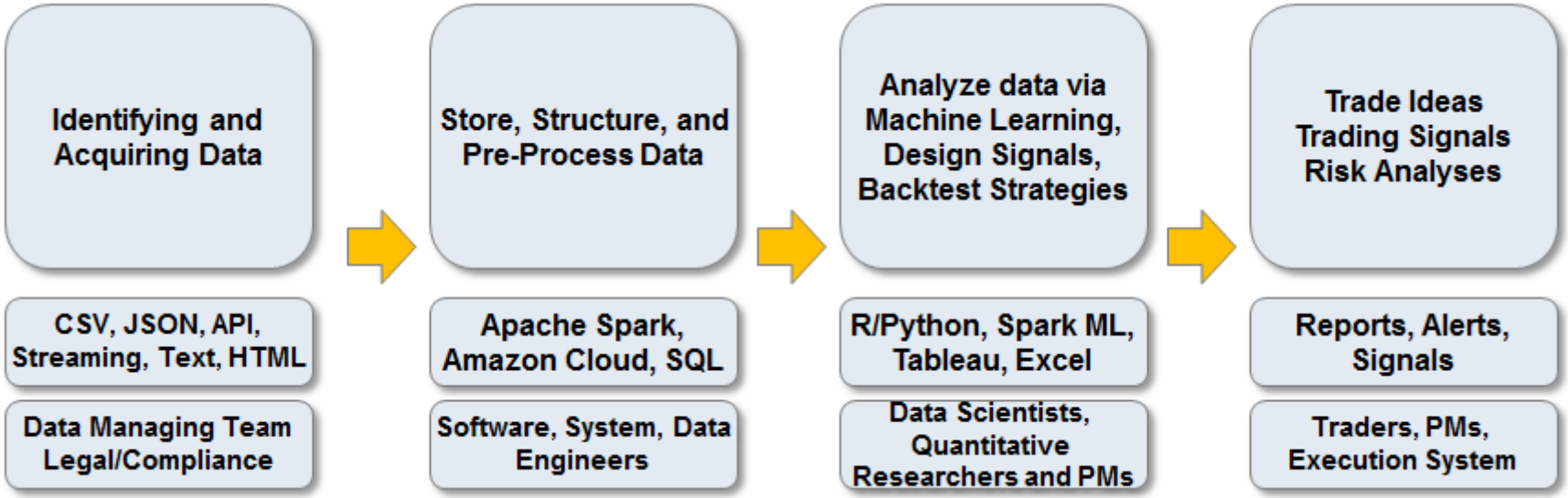
楊良河博士為香港大學統計及精算學系副教授

ML/AI Strategies



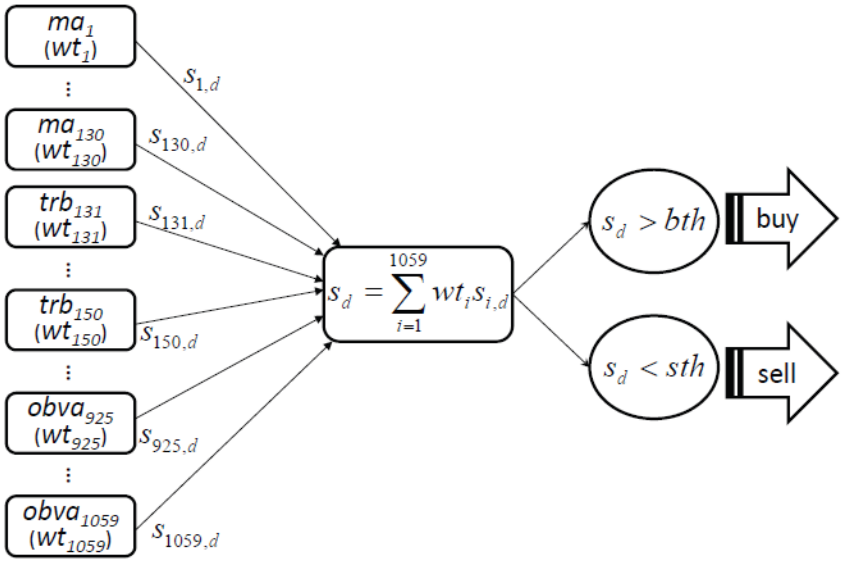
Source: J.P.Morgan Macro QDS

Developing a Trading Algorithm



Source: J.P.Morgan Macro QDS

Combining Technical Trading Rules:
Performance-based Reward Strategy (PRS)



for r_i ($profit_i < 0$) $\quad w_i = w_i - p$

for r_i ($profit_i > 0$) $\quad w_i = w_i + r$

where p and r depends on *rf*

- The weights of profitable/non-profitable rules should be increased/decreased
- Two time spans:
  - Memory span *ms*: length of the time period for evaluating the rule performance
  - Review span *rs*: how often to evaluate the performance and update the component rules' weights
- Reward factor *rf*: define how much for reward and penalty

技術分析與極速運算雙劍合璧

數裏見真章

王緋、楊良河

上星期筆者介紹了一種常用來處理大數據的極速運算技術 — Hadoop，本文會探討如何應用Hadoop來優化技術交易策略。大家都知道在金融市場上有林林總總的技術交易法則（technical trading rules），例如移動平均線（Moving Average），相對強弱指數（Relative Strength Index）等。到底哪一種技術交易法則最有效呢？答案是市場上沒有長勝將軍，沒有單一的交易法則能一直賺錢。

事實上，交易法則的有效性可以是輪替的，有效的法則可以因為太多人採納而變得失去威力，買入訊號出現時可以變成是沽出的好時機。就像女士的時裝潮流：此時流行短裙，彼時流行熱褲，掌握大數據者當然可以走在時代的前端，及早覺察出潮漲與潮退的時刻，技術分析者也希望早着先鞭，發掘出那個指標能成為明日之星。不過就像大數據一樣，類似的動態分析，如無極速運算的支援，難以成事。

整合逾千個技術交易法則

筆者最近就研究了一個整合大量交易法則的策略。我們先選取了7類最常用的技術交易法則，包括移動平均線（MA），相對強弱指數（RSI），交易區間突破（TRB），保力加通道（BB），隨機震盪指標（STO）、移動平均滙聚背馳指標（MACD）及能量潮指標（OBVA）。每一個交易法則都有一至四個參數（Parameters），例如移動平均線中就有兩個參數，分別代表計算短期和長期移動平均線的時段的天數。透過對每一類交易法則，考慮多種常見的參數組合，我們可得出總共1059個的基本交易法則。

假設每個基本交易法則每天會發出不同的信號：買入（1）、不變（0）或賣出（-1）。由於交易法則各有好壞，因此我們給每個交易法則一個權重（Weight）來表示它所發出的信號的可靠程度，並假設所有權重加起來是等於1。重要的是，權重可因時而變。

獎罰制調整權重

最終整合交易策略的信號，是所有基本交易法則所產生交易信號的加權平均值（Weighted Average）。得到的交易信號實際上仍然是一個量化的數值，我們將其與另外兩個控制買入和賣出的參數比較：若大於買入門檻值則買入股票，若小於賣出門檻值則出售股票，其餘就不變【圖1】。

任何交易法則的表現都會隨着時間的推移而時好時壞。因此我們會動態地根據每個基本交易法則的近期表現去調整它們的權重。簡單來說就是一個獎罰機制：對於那些近績優異的交易法則，就增加它們的權重；反之近績較差的法則，則降低權重。

PSO模擬群鳥覓食 尋最佳回報

根據以上描述，現在整個交易策略的參數，包括多達一千個的初始權重值以及獎罰制內的參數和買入／賣出門檻值等的參數。要找出最佳的參數從而達至最高的回報，我們採用了在1995年發明的粒子群優化法（Particle Swarm Optimization，PSO）來進行優化。簡單來說，這優化方法是模擬一群小鳥在尋找食物（即找尋最佳回報）：當一隻小鳥有新發現時（即回報有增長），它會通知其他小鳥這好消息。所以每隻小鳥每步的位置（即一組參數）會根據本身和同伴的飛行經驗來調整。

Hadoop助拳加速優化

有了策略，也有了優化方法，接下來的挑戰是更加高效快速地完成交易策略的優化。巨大的數據量加上多達一千個的子交易法則，讓快速優化看起來是不可能完成的任務。實際上我們在研究的最初階段，每次優化均要耗費好幾天的時間，嚴重地影響了研究進度。最終我們利用了大數據處理工具Hadoop，在一個小型的分散式運算集群上，實現了交易策略的快速優化，並把每次優化的時間縮減到兩小時左右。

我們以納斯特100（NASDAQ 100）股票指數中，由1995年至2002年的股票價格作為歷史數據，對我們的交易策略進行優化，並以2003年至2010年的股票價格，作為測試數據對優化後的交易策略進行驗證測試。

如【表】所示，在測試期間納斯特100市場本身的年回報率只是10.5%。

經過八年的測試交易，我們的交易策略（動態權重調整的交易策略）能達到21.8%左右的年回報率，即使是固定權重也有19%的年回報率，而所用到的一千多個基本交易法則，在測試期間達至最高的年回報率，也才是18.8%左右。

【圖2】展示了對交易策略裏，其中140個移動平均線交易法則，每個法則的權重在測試期間的變化，包括權重的最大值，最小值以及平均值（以中間橫線表示）。

由此可見，一些交易法則的變化很大，這正正反映市場的變化，一個好的交易策略須有自我更新的能力，我們的動態權重調整的交易策略，就是其中一個可行的方案。其他詳細結果，見【註】。

當然，如同簡單的量化交易策略一樣，複雜的交易策略也只是電腦根據數學模型進行的一種推斷而已。在真正的交易中，這些交易策略產生的信號，只是決定最後交易決策的一部分而已。例如投資者還會考慮各種最新的政經資訊，甚至社交人群的情緒波動等。但作為技術分析的一種手段，量化交易策略，也實實在在的給投資者提供了一些寶貴的市場訊息，發揮不可替代的作用。

註：Wang,F., Yu,P.L.H. and Cheung, D. W.（2014）. *Combining Technical Trading Rules Using Parallel Particle Swarm Optimization based on Hadoop*. Proceedings of 2014 IEEE World Congress on Computational Intelligence.

王緋為香港大學計算機科學系研究碩士畢業生，現任摩根士丹利投資銀行實習生

楊良河博士為香港大學統計及精算學系副教授

圖1 加權整合交易策略

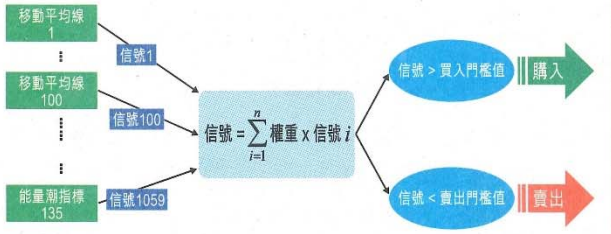


圖2 140個移動平均線交易法則在測試期權重變化

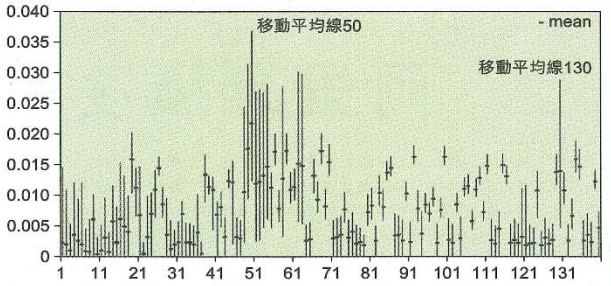
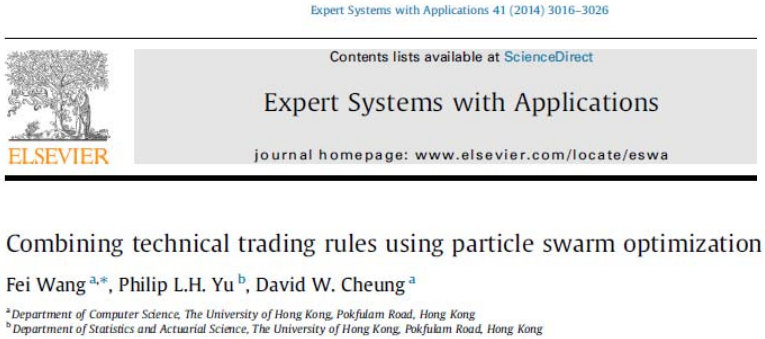


表 交易策略測試期間年回報率

交易策略／市場指數	測試期間年回報率
動態權重調整的交易策略	21.8%
固定權重的交易策略	19.0%
測試期間表現最佳的基本交易法則	18.8%
納斯特100股票指數	10.5%

Parallel Particle Swarm Optimization based on Hadoop


Expert Systems with Applications 41 (2014) 3016–3026

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Combining technical trading rules using particle swarm optimization

Fei Wang [a,*], Philip L.H. Yu [b], David W. Cheung [a]

[a] Department of Computer Science, The University of Hong Kong, Pokfulam Road, Hong Kong
[b] Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong

Combining Technical Trading Rules Using Parallel Particle Swarm Optimization based on Hadoop

Fei Wang, Philip L.H. Yu and David W. Cheung

Presented in WCCI 2014

- Data:
  - Training Period: 1995 - 2002, 8 years
  - Testing Period: 2003 - 2010, 8 years
  - NASDAQ100: 52 stocks enough recorded daily closed price
  - All data are from Reuters 3000Xtra

PRS Testing Performance

PERFORMANCE OF PRS AND THE SEVEN BEST COMPONENT RULES IN THE TESTING PERIOD

Trading rules (parameters)	ANP	Sharpe	Payoff	
PRS	21.8%	0.94	4.40	
MA (nl=150, ns=125)	18.8%	1.07	3.53	
TRB (n=125)	16.0%	1.05	5.84	
BBs (n=30, k=2.3)	18.6%	1.14	4.62	
RSI (n=13, ob=80, os=30)	9.0%	0.60	0.81	
STO (n=10, m=3, ob=90, os=20)	11.6%	0.76	0.74	14
MACD (nl=100, ns=40, m=15)	11.4%	0.82	2.75	16
OBVA (nl=75, ns=50)	10.7%	0.72	1.49	2

STOCK PROFIT SUMMARY OF PRS AND THE SEVEN BEST COMPONENT RULES IN THE TESTING PERIOD

Trading rules (Parameters)	Summary	Profitable stocks[a]	Non-profitable stocks[a]	All stocks (Profit ratio)[b]
PRS	No.Stocks[c]	44	8	52(84.6%)
	Net Profit	20.22	-0.27	19.95
MA (150, 125)	No.Stocks	34	18	52(65.4%)
	Net Profit	15.89	-0.45	15.44
TRB (125)	No.Stocks	41	11	52(78.8%)
	Net Profit	12.15	-0.25	11.90
BBs (30, 2.3)	No.Stocks	39	13	52(75.0%)
	Net Profit	15.44	-0.33	15.11
RSI (13, 80, 30)	No.Stocks	43	9	52(82.7%)
	Net Profit	5.40	-0.26	5.14
STO (10, 3, 90, 20)	No.Stocks	46	6	52(88.5%)
	Net Profit	7.47	-0.16	7.31
MACD (100, 40, 15)	No.Stocks	37	15	52(71.2%)
	Net Profit	7.51	-0.40	7.11
OBVA (75, 50)	No.Stocks	38	14	52(73.1%)
	Net Profit	6.75	-0.24	6.51

[a] Profitable stock means the stock whose final equity is more than its initial equity and vice versa.

[b] Profit ratio = (number of profitable stocks) / (total number of stocks).

[c] No.Stocks is the number of stocks. Net Profit is in million dollars.

Yearly Performance

	PRS	MA (150, 125)	TRB (125)	BBs (30, 2.3)	RSI (13, 80, 30)	STO (10, 3, 90, 20)	MACD (100, 40, 15)	OBVA (75, 50)
2003	**63.4752%**[1]	60.6130%	42.8268%	46.1083%	29.7601%	42.0463%	34.4278%	24.9238%
2004	**27.1650%**	15.4416%	18.4555%	19.8893%	10.6805%	12.1459%	5.1181%	18.8698%
2005	**15.2504%**	14.6832%	10.0164%	8.5618%	11.5607%	10.6413%	8.3183%	7.9970%
2006	20.4195%	7.5612%	12.0910%	**28.0389%**	4.4490%	8.0529%	21.0010%	6.8026%
2007	**33.3180%**	27.5845%	25.0693%	20.3132%	7.0426%	6.0613%	5.2955%	5.6800%
2008	-33.1956%	-11.3090%	-18.2505%	-4.8854%	-16.6836%	-16.7566%	**2.0170%**	-8.5392%
2009	**56.1281%**	36.0397%	35.3156%	34.4213%	20.9406%	33.4651%	13.8014%	26.8058%
2010	**21.0858%**	14.6616%	16.5620%	6.6916%	11.4287%	8.3006%	4.1748%	10.8166%
Average	**25.4558%**	20.6595%	17.7608%	19.8924%	9.8973%	12.9946%	11.7692%	11.6696%
Std.Dev[2]	27.4763%	19.9162%	17.2509%	15.3358%	12.5827%	16.7666%	**10.3296%**	10.8328%

[1] The highest profit for each year is in bold type.
[2] Std.Dev means the standard deviation of 8 years' profits.

文本資訊蘊藏投資先機

數裏見真章

李青龍、楊良河

今年初，畢非德宣布旗下的通訊社Business Wire決定不再向高頻交易商，提供特許即時新聞訊息傳送的付費服務。事緣是因一篇《華爾街日報》的報道，指Business Wire為一些高頻交易者提供即時市場動向消息，讓交易者可以比利用其他供應商如Thomson Reuters讀取新聞的用戶早一瞬間（如一秒）得到消息而獲利。

Business Wire為保公司聲譽，決定與高頻交易商劃清界線。雖然只是一秒鐘或更短的時間，對高頻交易商來說，已能從中及時獲取資訊，並根據資訊作出正確的決策而獲利和避免不必要的損失。

文字為本市場數據

一般量化投資策略使用的都是結構化的市場數據，如每日股票最高價、最低價、成交量等，但金融資訊中絕大部分是以文本形式存在的非結構化數據，例如財經新聞、股票分析報告和股票評論等媒介資訊。它們都能迅速地反映市場最新動態和投資者的情緒。相信讀者們也經常閱讀財經新聞、關注公司動向，如果我們能有效地挖掘出這裏面的蘊藏着的資訊，我們就可以從一個全新的角度建立投資策略。現今的高頻交易商不僅處理及分析價格數據速度快，它們連處理新聞資訊也快如閃電，買與賣都能早着先鞭，自然能够所向披靡。

新浪微博與股市何干？

Bollen， Mao和Zeng在2011年發表了一篇名為*Twitter mood predicts the stock market*的文章，他們通過搜集Twitter上的文本數據，利用文本挖掘技術來評估大眾的情緒從而進行股票價格預測。筆者也在此湊湊熱鬧，利用分析新浪微博的短訊，來看看能否預測滬深300指數的走勢。

新浪微博是一款類似Twitter的社交網站，是中國大陸用戶最多的微博平台網站。筆者收集了144位活躍寫投資評論的股票分析師用戶所發布的歷史微博短訊。由於新浪微博允許用戶用若干標籤(tag word)來描述自己的身份，我們統計了一下這144位分析師所使用的標籤，按照頻率從高到低排列得出【圖1】。

我們發現，這些分析師除了會使用一些金融用詞「股票」、「分析師」等來描述自己外，也會像大多數人一樣使用諸如「電影」、「美食」等標籤。這些標籤沒法顯示出金融分析師這一群體的特點，於是我們可以運用文本挖掘中常用的TF-IDF技術。

TF-IDF是何物？

TF-IDF(term frequency-inverse document frequency) 是一種用於評估一字詞對於一個文件集的重要程度。實際上，TF-IDF是TF x IDF，其中TF表示每個字詞在每個文件中出現的次數。【圖1】展示了各個標籤在所有用戶的短訊中出現的次數。若某字詞出現次數愈多，該字詞應愈重要。另外，IDF叫做逆向文件頻率，當含有一字詞的文件愈多，這個字詞的IDF就愈小，這亦表示這個字詞用來區分不同文件的能力不強。這樣TF-IDF兼顧了字詞的出現頻率與字詞的代表性。例如「分析師」在144的用戶中出現了15次。利用一個擁有15149652相關的微博用戶庫中，「分析師」一詞在1025用戶的標籤出現過，其IDF就是1+log2(15149652/1025)=14.851，最後「分析師」的TF-IDF值是15 x 14.851 = 222.77。

詞頻與向量空間模型

當使用TF-IDF對用戶的標籤進行統計的時候，「電影」、「美食」這種日常高頻詞彙由於IDF值很小，因此TF-IDF權重不是很高。我們可以看出利用TF-IDF篩選出的標籤，更好地描述了金融分析師這個群體【圖2】。

既然說要通過文本資訊來開發量化投資策略，那麼首先我們要做的就是將文本量化。在上星期文章【註】中，我們介紹了中文的分詞。當我們將一段話中的每一句話都分詞之後，我們就可以統計每一個詞出現的頻率。比如本文的第二段話，出現頻率最高的5個詞是資訊(4次)，新聞(3次)，投資(3次)，數據(3次)及股票(3次)。我們將出現頻率看成一個向量(4，3，3，3，3)，用來表示一段文字的特徵，也就是說我們成功的將一段文字量化成為了一個向量，這就是向量空間模型(Vector Space Model)，它是文本挖掘中最常用的模型之一。

除了使用頻率之外，我們也可以使用其他權重來構建一個文件的向量，比如我們剛剛所提到的TF-IDF。成功地將文本量化成為向量之後，我們就可以使用傳統方法來進行建模了。筆者將144位用戶，在2011年7月至2012年12月期間，每日所發的微博文本，利用TF-IDF量化成為空間向量並進行適度的降維(dimension reduction)，配合我們常用的歷史價格數據及技術分析指標(如Williams %R指數)，使用Logistic回歸對滬深300指數在未來5日、10日和15日內是漲是跌進行預測。得到的結果如下【表】所示。

我們可以看出，當添加了文本資訊到預測模型中後，可以有效的提高對股市預測的準確性。

這說明了用戶在社交網絡上所發的股評內容包含了預測股市的有用資訊，運用得宜時，便能在現今千變萬化的市場得着先機。

註：《數裏見真章》2014年5月22日「本來無數字，何處來分析？」

李青龍為香港大學統計學項士畢業生

楊良河博士為香港大學統計及精算學系副教授

■社交網絡上所發的股評內容包含了預測股市的有用資訊，如運用得宜，能在市場得着先機。（資料圖片）

圖1 新浪微博144位分析師使用標籤頻率

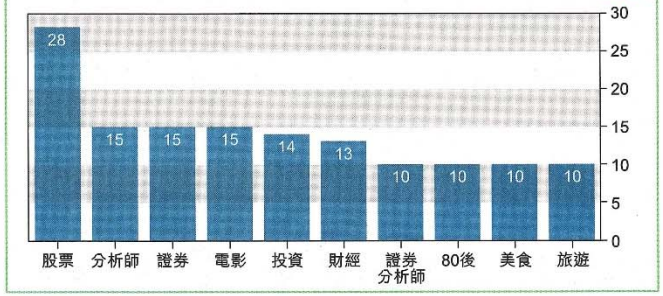
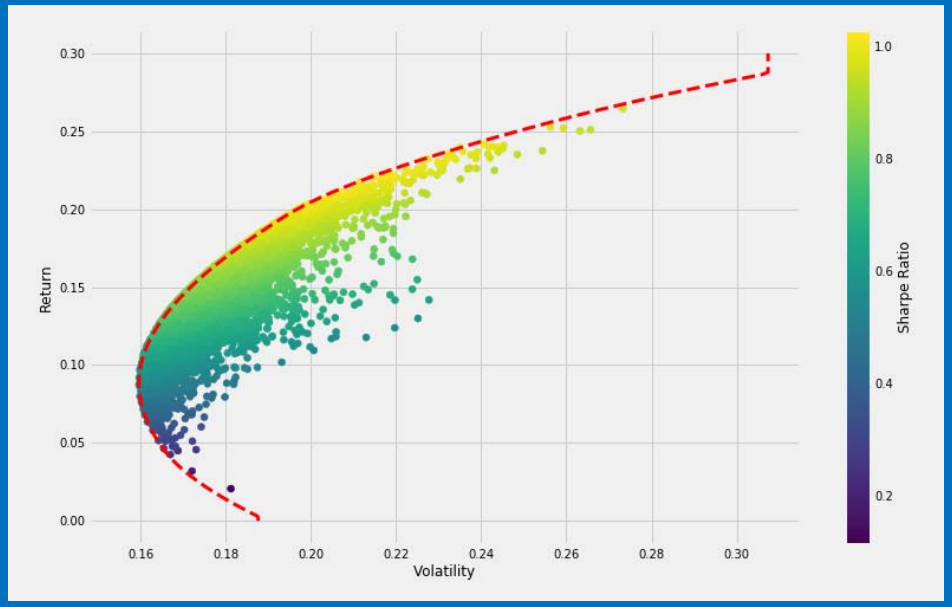


圖2 新浪微博144位用戶使用標籤TF-IDF值

Logistic回歸預測錯誤率
（測試期間為2013年1月至3月）

	5日的升跌	10日的升跌	15日的升跌
不利用文本資訊	44.1%	38.9%	32.2%
利用文本資訊	37.3%	13.6%	16.9%

Portfolio Optimization



Source: https://towardsdatascience.com/python-markowitz-optimization-b5e1623060f5

Markowitz Mean-Variance Portfolio Optimization

- Investors usually prefer high return and low risk.
- The optimal portfolio can be determined by maximizing the utility function:

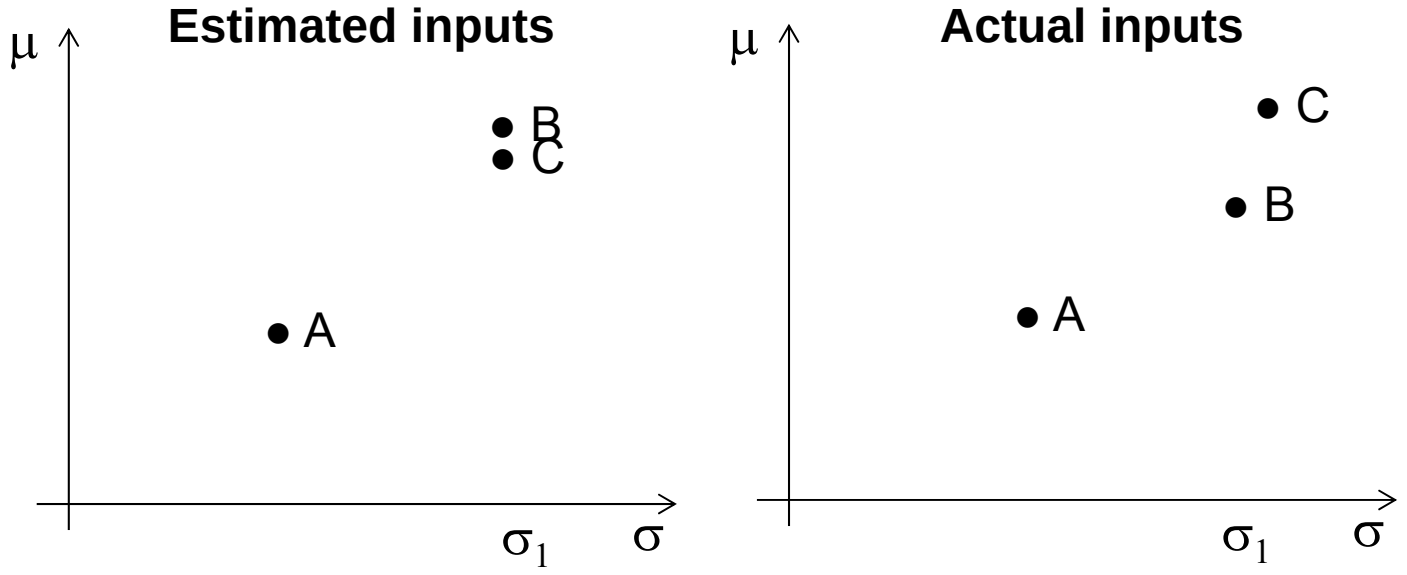
$$\boldsymbol{w} = \arg\max_{\boldsymbol{w}} E(R) - \frac{\gamma}{2} Var(R)$$

subject to $\boldsymbol{w}'\mathbf{1} = 1$ and other constraints

- $E(R) = \boldsymbol{w}'\boldsymbol{\mu}$
- $Var(R) = \boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w}$
- $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the mean vector and covariance matrix of rate of returns on k risky assets

Markowitz Optimization Enigma

- In practice, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are unknown. A common approach is to plug in historical sample mean and covariance matrix.
- Simply plugging in their estimates into the optimization will then ignore the estimation error.
- **Example:** Suppose you can tolerate the portfolio risk to be at most σ_1. What does the optimal portfolio look like?



Solutions…

- Binding constraints on the weights (Best, et al. 1991)
- Resampled efficient frontier (Michaud, 1998)
- Shrinkage estimation of the weights (Golosnoy and Okhrin, 2007)

- However, these works focused on the point estimation of the optimal portfolio weights.
  - Yu, et al. (2017a): constructed confidence intervals for the optimal portfolio weights based on Generalized Pivotal Quantity
  - Yu, et al. (2017b): considered a penalized approach for high-dimensional portfolio optimization

Generalized Pivotal Quantity (GPQ)

- Notations:
  - $\boldsymbol{X}$: a random vector
  - $\boldsymbol{X}^*$: an independent copy of $\boldsymbol{X}$
  - $\boldsymbol{x}$: observed value of $\boldsymbol{X}$

- $\widetilde{\boldsymbol{w}}(\boldsymbol{X}, \boldsymbol{X}^*, \boldsymbol{w})$ is a GPQ of $\boldsymbol{w}$ if
  - $\widetilde{\boldsymbol{w}}(\boldsymbol{x}, \boldsymbol{x}, \boldsymbol{w})$ only depends on $\boldsymbol{w}$
  - $\widetilde{\boldsymbol{w}}(\boldsymbol{X}, \boldsymbol{x}, \boldsymbol{w})$ has a probability distribution free of unknown parameters

- Confidence intervals for $\boldsymbol{w}$:
  - Construct $\boldsymbol{C}_\alpha$ so that $\mathrm{P}(\widetilde{\boldsymbol{w}}(\boldsymbol{X}, \boldsymbol{x}, \boldsymbol{w}) \in \boldsymbol{C}_\alpha) = 1 - \alpha$

Generalized Pivotal Quantity (GPQ)

- Useful when nuisance parameter exists or classic method fails
  - Mean of lognormal distribution (Krishnamoorthy, Mathew 2003)
  - MANOVA and mixed models (Weerahandi 2004)
  - Correlation coefficient (Krishnamoorthy and Xia 2007)

- Has good property: if $\tilde{\theta}$ is a GPQ for θ, $\mathrm{f}(\tilde{\theta})$ is a GPQ for $\mathrm{f}(\theta)$ for any real-valued function.

- However,
  - there is no theoretical guarantee that in small sample size scenarios, a generalized confidence interval will provide coverage probabilities close to the nominal level. In fact the coverage could depend on nuisance parameters. Simulation study is necessary to measure its performance.

Example: GPQ for $\boldsymbol{\Sigma}$

- $\boldsymbol{X} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$, where $\boldsymbol{X}_i \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- Then $\boldsymbol{A} = \sum_{i=1}^{n}(\boldsymbol{X}_i - \overline{\boldsymbol{X}})(\boldsymbol{X}_i - \overline{\boldsymbol{X}})' \sim W_k(\boldsymbol{\Sigma}, n-1)$
- Write $\boldsymbol{A} = \boldsymbol{T}\boldsymbol{T}'$ and $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}'$, $\boldsymbol{T}$ and $\boldsymbol{\Gamma}$: lower triangular matrices
- Denote $\boldsymbol{B} = \boldsymbol{\Gamma}^{-1}\boldsymbol{T}$ and we have:

$$\boldsymbol{B}\boldsymbol{B}' = \boldsymbol{\Gamma}^{-1}\boldsymbol{T}\boldsymbol{T}'\boldsymbol{\Gamma}'^{-1} \sim W_k(\boldsymbol{I}, n-1)$$

- GPQ for $\boldsymbol{\Gamma}$ is

$$\widetilde{\boldsymbol{\Gamma}}(\boldsymbol{X}, \boldsymbol{X}^*, \boldsymbol{\Gamma}) = \boldsymbol{T}^*(\boldsymbol{\Gamma}^{-1}\boldsymbol{T})^{-1} = \boldsymbol{T}^*\boldsymbol{B}^{-1}$$

  - $\widetilde{\boldsymbol{\Gamma}}(\boldsymbol{x}, \boldsymbol{x}, \boldsymbol{\Gamma}) = \boldsymbol{\Gamma}$
  - $\widetilde{\boldsymbol{\Gamma}}(\boldsymbol{X}, \boldsymbol{x}, \boldsymbol{\Gamma}) = \boldsymbol{T}_0\boldsymbol{B}^{-1}$ is distribution free, where $\boldsymbol{T}_0$ is a realized value of $\boldsymbol{T}$

- If $\tilde{\theta}$ is a GPQ for θ, $\mathrm{f}(\tilde{\theta})$ is a GPQ for $\mathrm{f}(\theta)$ for any real-valued function. Then GPQ for $\boldsymbol{\Sigma}$ is:

$$\widetilde{\boldsymbol{\Sigma}}(\boldsymbol{X}, \boldsymbol{X}^*, \boldsymbol{\Sigma}) = \widetilde{\boldsymbol{\Gamma}}\widetilde{\boldsymbol{\Gamma}}' = \boldsymbol{T}^*\boldsymbol{B}^{-1}(\boldsymbol{T}^*\boldsymbol{B}^{-1})'$$

GPQ for $\boldsymbol{\mu}$

- $\boldsymbol{X} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$, where $\boldsymbol{X}_i \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- Let $\overline{\boldsymbol{X}}$ be the sample mean. Then

$$\boldsymbol{Z} = \sqrt{n}\boldsymbol{\Gamma}^{-1}(\overline{\boldsymbol{X}} - \boldsymbol{\mu}) \sim N_k(\boldsymbol{0}, \boldsymbol{I})$$

- GPQ for $\boldsymbol{\mu}$ is

$$\tilde{\boldsymbol{\mu}}(\boldsymbol{X}, \boldsymbol{X}^*, \boldsymbol{\mu}) = \overline{\boldsymbol{X}}^* - \frac{1}{\sqrt{n}}\tilde{\boldsymbol{\Gamma}}\sqrt{n}\boldsymbol{\Gamma}^{-1}(\overline{\boldsymbol{X}} - \boldsymbol{\mu}) = \overline{\boldsymbol{X}}^* - \frac{1}{\sqrt{n}}\tilde{\boldsymbol{\Gamma}}\boldsymbol{Z}$$

  - $\tilde{\boldsymbol{\mu}}(\boldsymbol{x}, \boldsymbol{x}, \boldsymbol{\mu}) = \boldsymbol{\mu}$
  - $\tilde{\boldsymbol{\mu}}(\boldsymbol{X}, \boldsymbol{x}, \boldsymbol{\mu}) = \overline{\boldsymbol{x}} - \frac{1}{\sqrt{n}}\tilde{\boldsymbol{\Gamma}}\boldsymbol{Z}$ is distribution free, where $\overline{\boldsymbol{x}}$ is a realized value of $\overline{\boldsymbol{X}}$

GPQ for $\boldsymbol{w}$

- GPQ for $\boldsymbol{w}$ is

$$\tilde{\boldsymbol{w}} = \arg\max_{\tilde{\boldsymbol{w}}} \tilde{\boldsymbol{w}}'\tilde{\boldsymbol{\mu}} - \frac{\gamma}{2}\tilde{\boldsymbol{w}}'\tilde{\boldsymbol{\Sigma}}\tilde{\boldsymbol{w}}$$

  subject to $\tilde{\boldsymbol{w}}'\mathbf{1} = 1$ and other constraints

- Check the two requirements of GPQ:
  1. The observation of $\tilde{\boldsymbol{w}}$ is the true $\boldsymbol{w}$ since the observations of $\tilde{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{\Sigma}}$ are the true $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.
  2. Conditional on $\boldsymbol{X}^* = \boldsymbol{x}$, both $\tilde{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{\Sigma}}$ do not depend on any parameter, so $\tilde{\boldsymbol{w}}$ is also distributional free.
- MC simulation can be used to estimate the distribution of $\tilde{\boldsymbol{w}}$
  - Point estimator: the average of all simulated $\tilde{\boldsymbol{w}}$
  - Interval estimator: $100(1-\alpha)\%$ generalized confidence interval $\boldsymbol{C}_\alpha$ is a set satisfying $\mathrm{P}(\tilde{\boldsymbol{w}} \in \boldsymbol{C}_\alpha) = 1-\alpha$

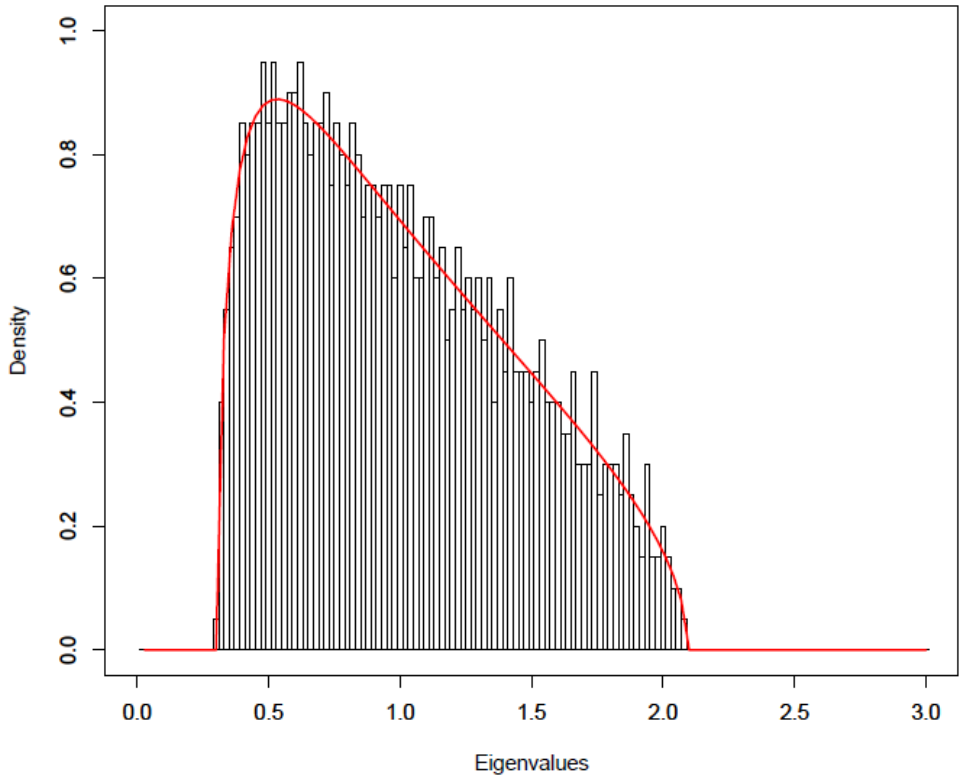
Application: Portfolio Rebalancing

- Eight NYSE stocks: XOM, GE, IBM, T, PG, JNJ, JPM, KO
- Monthly data from June 1986 to May 2011 (25 years)
- Training: 10-year data
- Transaction cost: 0.05% per trade
- Monthly rebalancing vs Dynamic rebalancing

γ	Measure	Plug-in	Resampled	GPQ	GPQ-95	GPQ-90
	Final Wealth	3.8671	3.9121	3.9878	4.0215	4.0297
40	Sharpe Ratio	2.4117	2.4309	2.4579	2.4605	2.4642
	# updating	180	180	180	163	163
	Final Wealth	3.6234	3.6749	3.7613	3.8258	3.8213
25	Sharpe Ratio	2.3164	2.3378	2.3732	2.3869	2.3821
	# updating	180	180	180	141	148
	Final Wealth	2.7704	3.0403	3.1295	3.1229	3.0260
10	Sharpe Ratio	1.8502	1.9944	2.0469	2.0117	1.9714
	# updating	180	180	180	27	40

Curse of Dimensionality

- Estimating high-dimensional covariance matrices is challenging.
- Consider $\boldsymbol{x}_t = (x_{1t}, \ldots, x_{nt})' \sim N_n(\boldsymbol{0}, \boldsymbol{I}_n)$. Then the density of eigenvalues of the empirical covariance matrix $\boldsymbol{S} = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}_t \boldsymbol{x}_t{}'$



Marčenko-Pastur density
T = 5000
n = 1000

Penalized Likelihood Method

- Matrix logarithm: $\boldsymbol{A} = \log \boldsymbol{\Sigma}$
- The negative log-likelihood function is

$$L(\boldsymbol{A}) = tr(\boldsymbol{A}) + tr[\exp(-\boldsymbol{A})\,\boldsymbol{S}]$$

- Deng and Tsui (2013) considered the penalty function:

$$\|\boldsymbol{A}\|_F^2 = tr(\boldsymbol{A}^2) = \sum_{i=1}^{n} [\log d_i]^2$$

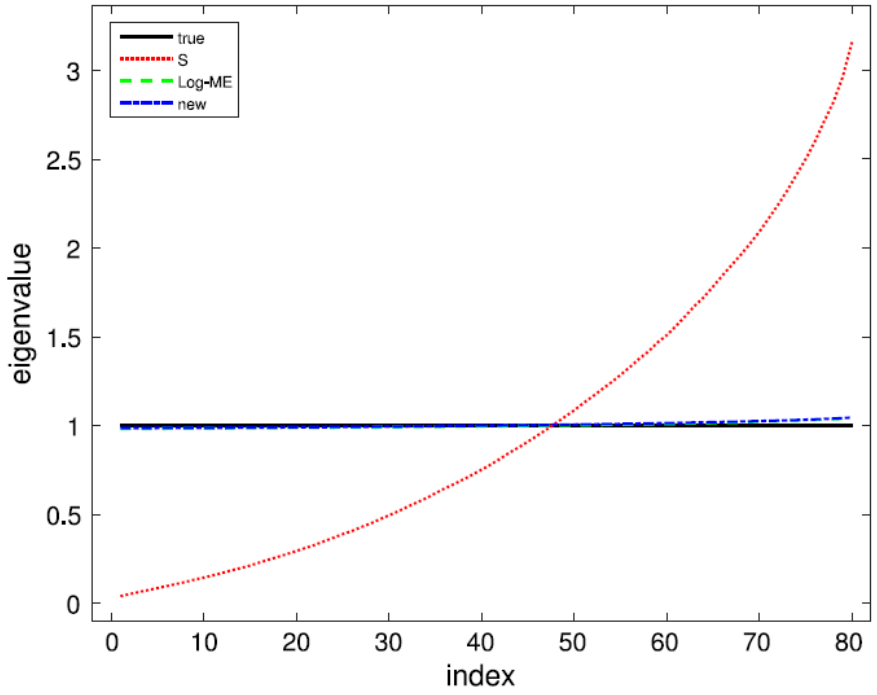
where $d_1, \ldots, d_n$ are the eigenvalues of $\boldsymbol{\Sigma}$

- Yu, et al. (2017) considered an alternative penalty function:

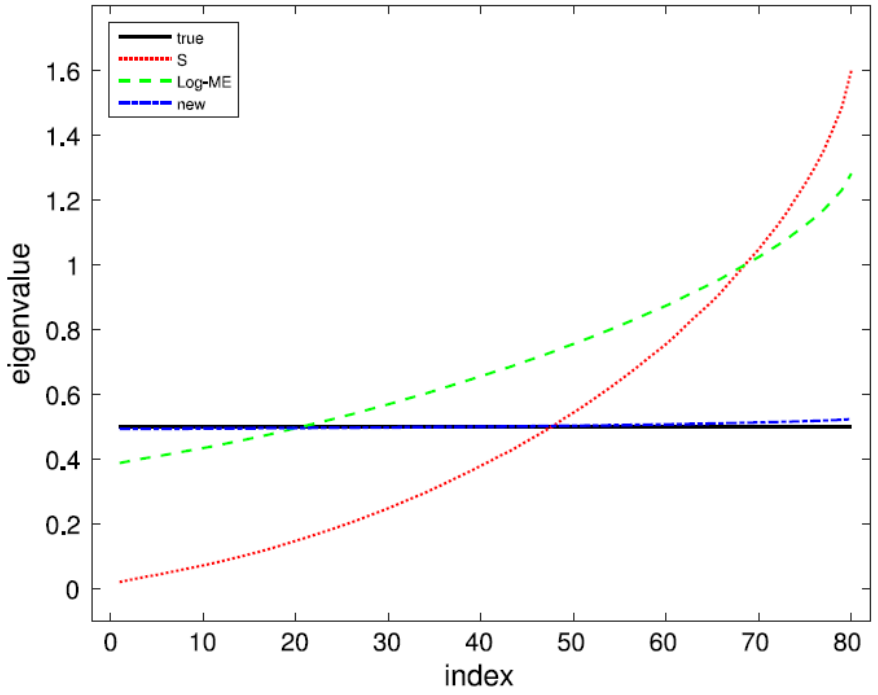
$$\|\boldsymbol{A} - m\boldsymbol{I}_n\|_F^2 = \sum_{i=1}^{n} [\log d_i - m]^2$$

where m is an estimate of the average eigenvalue of $\boldsymbol{A}$

Why New Penalty



(a) True eigenvalue $= 1$.



(b) True eigenvalue $= 0.5$.

- S = sample covariance matrix
- log-ME = Deng and Tsui (2013)
- New = our method

Portfolio Optimization of S&P100 Stocks

- Data: T = 321 weekly returns of n = 93 constituent stocks of S&P100 index (7 stocks were unlisted after January, 2009), from the first week of 2009 to the 9th week of 2015.
- Use the first 135 weeks as training set, the next 135 weeks as validation set, and the last 51 weeks as testing set.

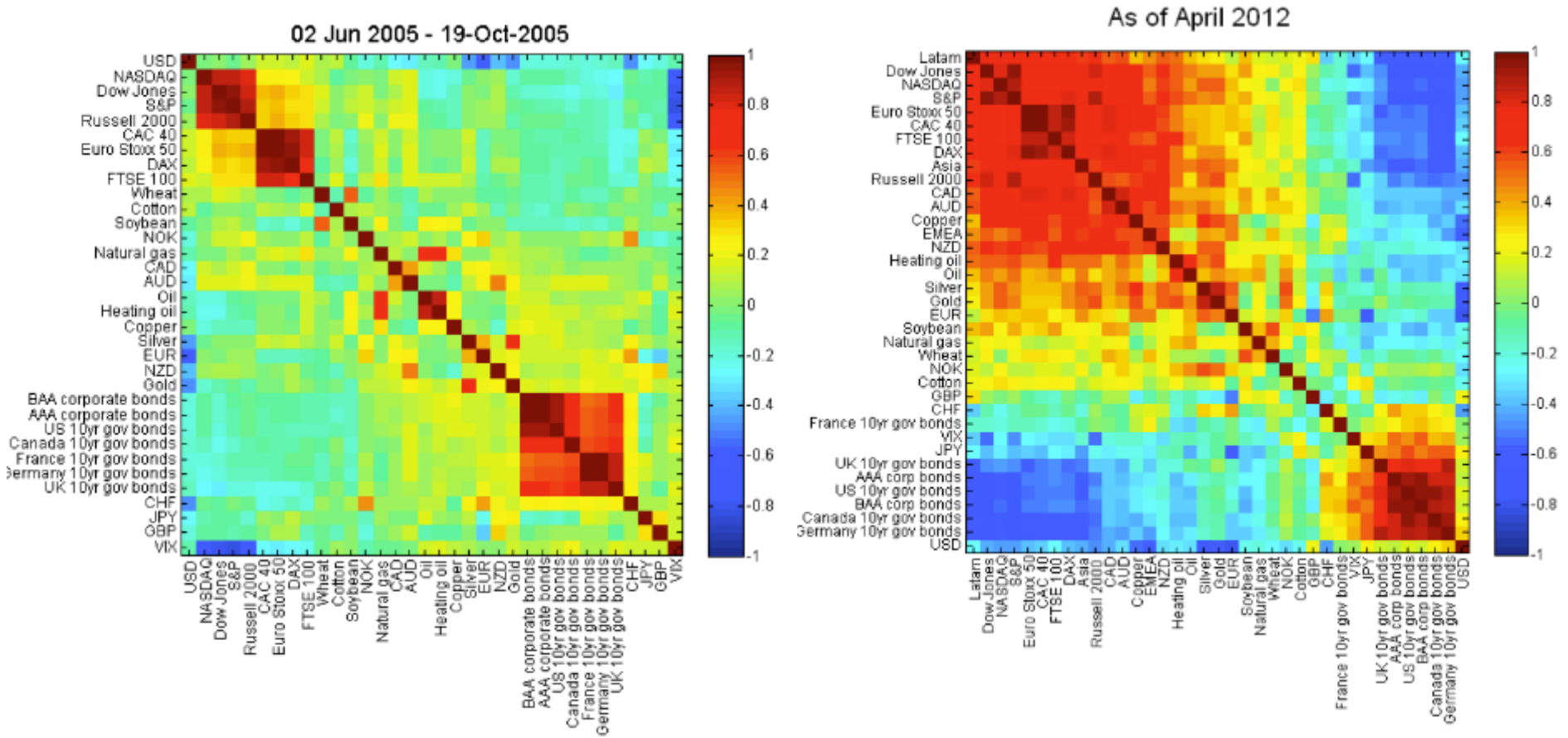
The realized return and realized risk of Global minimum variance portfolio (Problem (A)) in the testing period.

	S	Log-ME	new	Log-ME(S)	new(S)
Realized Return	0.114	0.157	0.162	0.152	0.164
Realized Risk	0.017	0.014	0.013	0.014	0.012

The realized return and realized risk of the optimal portfolio for Problem (B) in the testing period.

		S	Log-ME	new	Log-ME(S)	new(S)
$\sigma_p = 0.15/\sqrt{52}$	Realized Return	0.024	0.133	0.149	0.110	0.167
	Realized Risk	0.063	0.023	0.022	0.035	0.025
$\sigma_p = 0.20/\sqrt{52}$	Realized Return	−0.009	0.121	0.142	0.093	0.169
	Realized Risk	0.085	0.032	0.029	0.047	0.033
$\sigma_p = 0.25/\sqrt{52}$	Realized Return	−0.042	0.111	0.135	0.077	0.170
	Realized Risk	0.106	0.040	0.036	0.059	0.042

Dynamic Correlation Matrix



The availability of high-frequency intraday financial data enables us to estimate daily covariance matrix of asset returns directly, which leads to the so-called realized covariance (RCOV) matrix.

Existing Models for RCOV Matrices

- Wishart AR (WAR(r)) model (Gourieroux et al., 2009)

$$\boldsymbol{Y}_t \mid \mathcal{F}_{t-1} \sim W_n(\nu, \boldsymbol{\Lambda}_t, \boldsymbol{\Sigma}),$$

$$\boldsymbol{\Lambda}_t = \sum_{k=1}^{r} \boldsymbol{M}_k \boldsymbol{Y}_{t-k} \boldsymbol{M}_k'. \tag{1}$$

- Conditional AR Wishart (CAW(p, q)) model (Golosnoy et al., 2012)

$$\boldsymbol{Y}_t \mid \mathcal{F}_{t-1} \sim W_n(\nu, 0, \boldsymbol{\Sigma}_t/\nu),$$

$$\boldsymbol{\Sigma}_t = \boldsymbol{C}\boldsymbol{C}' + \sum_{i=1}^{p} \boldsymbol{B}_i \boldsymbol{\Sigma}_{t-i} \boldsymbol{B}_i' + \sum_{j=1}^{q} \boldsymbol{A}_j \boldsymbol{Y}_{t-j} \boldsymbol{A}_j'. \tag{2}$$

- Generalized CAW (GCAW(p, q, r)) Model (Yu, et al. 2017)

$$\boldsymbol{Y}_t \mid \mathcal{F}_{t-1} \sim W_n(\nu, \boldsymbol{\Lambda}_t, \boldsymbol{\Sigma}_t),$$

$\boldsymbol{\Lambda}_t$ is from (1) and $\boldsymbol{\Sigma}_t$ is from (2).

- However, fitting these models will be computationally demanding for moderate and high dimensions (say $n > 10$).

Existing Models for RCOV Matrices

- Matrix Factor Analysis (MFA) model (Tao et al. 2011)

$$\boldsymbol{Y}_t = \boldsymbol{L}\boldsymbol{F}_t\boldsymbol{L}' + \boldsymbol{E}_0,$$

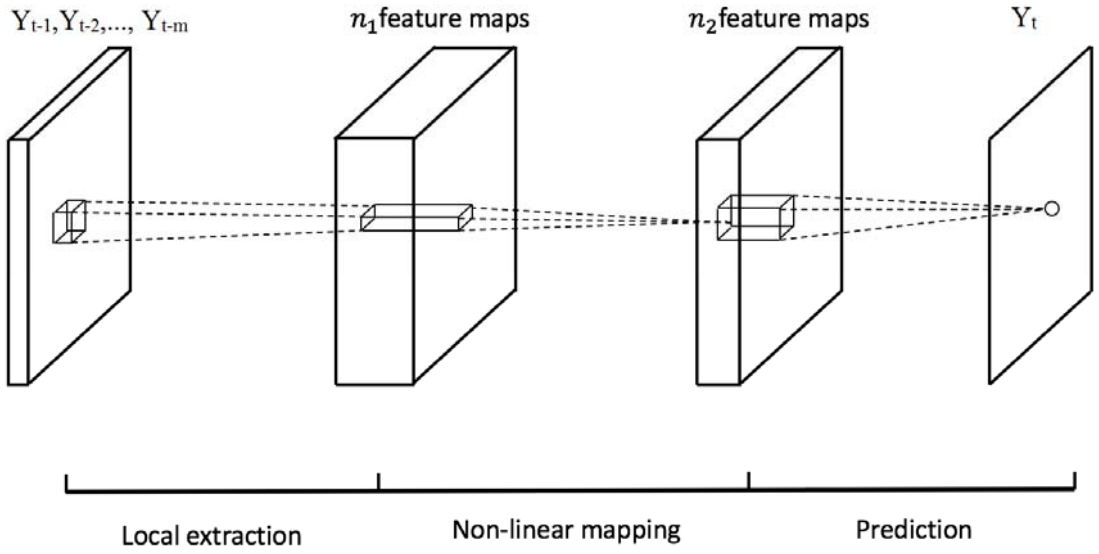
  where $\boldsymbol{L}$ is a $n \times d$ factor loading matrix, $\boldsymbol{F}_t$ are $d \times d$ positive definite matrices and $\boldsymbol{E}_0$ is a $n \times n$ positive definite constant matrix.
- To forecast $\boldsymbol{Y}_{t+1}$, Tao et al. (2011) adopted a two-step procedure by first estimating $\boldsymbol{L}$ and $\boldsymbol{F}_t$ and then fitting a VAR model to $vech(\hat{\boldsymbol{F}}_t)$:

$$\text{vech}(\hat{\boldsymbol{F}}_t) = \boldsymbol{\lambda}_0 + \sum_{j=1}^{q} \boldsymbol{\Lambda}_j \text{vech}(\hat{\boldsymbol{F}}_{t-j}) + \boldsymbol{e}_t$$

- Shen et al. (2015) replaced the VAR model by a diagonal CAW (DCAW) model with diagonal matrices for $\boldsymbol{A}, \boldsymbol{B}$ and $\boldsymbol{C}$.
- However, the loading matrix $\boldsymbol{L}$ is assumed to be constant over time, meaning that the dynamic correlation structure of $\boldsymbol{Y}_t$ is completely governed by the model (VAR or DCAW) for the low-dimensional $\hat{\boldsymbol{F}}_t$.

Deep Learning for RCOV Matrices

- The input is a set of historical RCOV matrices: a 3D cube $\widetilde{\boldsymbol{Y}}_{t-1} = (\boldsymbol{Y}_{t-1}, \boldsymbol{Y}_{t-2}, \ldots, \boldsymbol{Y}_{t-m})$ with dimension $m \times n \times n$, where m is the lag length and n is the number of assets.
- Our objective is to produce one-day ahead forecast of the RCOV matrix, $\boldsymbol{Y}_t$.
- We wish to learn a mapping F to connect $\widetilde{\boldsymbol{Y}}_{t-1}$ and $\boldsymbol{Y}_t$, which could handle high dimensional covariance matrices.
- Our mapping F is designed to consist of at least three convolutional layers (i.e., no pooling layers, no fully-connected layers):



Applications

- **3 datasets:**
  1. 25 constituent stocks of the DJIA index (18 Jan 2007 to 31 Dec 2013)
  2. 60 constituent stocks of the S&P100 index (10 Sep 2003 to 30 Dec 2016)
  3. 244 constituent stocks of the S&P500 index (10 Sep 2003 to 30 Dec 2016)
- High frequency data are downloaded from the NYSE TAQ database, and daily RCOV matrices are calculated using the averaging realized volatility matrix (ARVM) method proposed by Wang et al. (2010).
- Stock trading is from 9:30 am to 4:00 pm each day, with observations before 10:00 am deleted to avoid opening effects. Stocks with less than 100 daily trading records are also deleted.
- The testing set contains the last 252 days' RCOV matrices, the validation set contains the second last 252 days' RCOV matrices, the remaining matrices all go to the training set.

Forecasting Performance based on RMSE of the Testing Set

Models	DJIA	S&P100	S&P500
Moving Average	3.851 (7)	17.430 (3)	110.689 (5)
Exponential Moving Average	3.734 (7)	16.277 (5)	107.586 (6)
MFA-VAR	4.380	18.639	140.490
MFA-DCAW	4.246	16.916	115.144
Our model	3.367	14.914	103.044

Numbers in the brackets are best lag orders used for moving average and exponential moving average.

Matrix Factor Analysis (MFA): $\boldsymbol{Y}_t = \boldsymbol{L}\boldsymbol{F}_t\boldsymbol{L}' + \boldsymbol{E}_0$

Vector Autoregressive (VAR): $\mathrm{vech}(\boldsymbol{F}_t) \sim$ VAR(1)

Diagonal Conditional AR Wishart (DCAW): $\boldsymbol{F}_t \sim \boldsymbol{W}(v, \boldsymbol{\Sigma}_t/v)$ where

$\boldsymbol{\Sigma}_t = \boldsymbol{C}\boldsymbol{C}' + \boldsymbol{B}\boldsymbol{\Sigma}_{t-1}\boldsymbol{B}' + \boldsymbol{A}\boldsymbol{Y}_{t-1}\boldsymbol{A}'$, $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ are diagonal matrices

Fintech Challenges

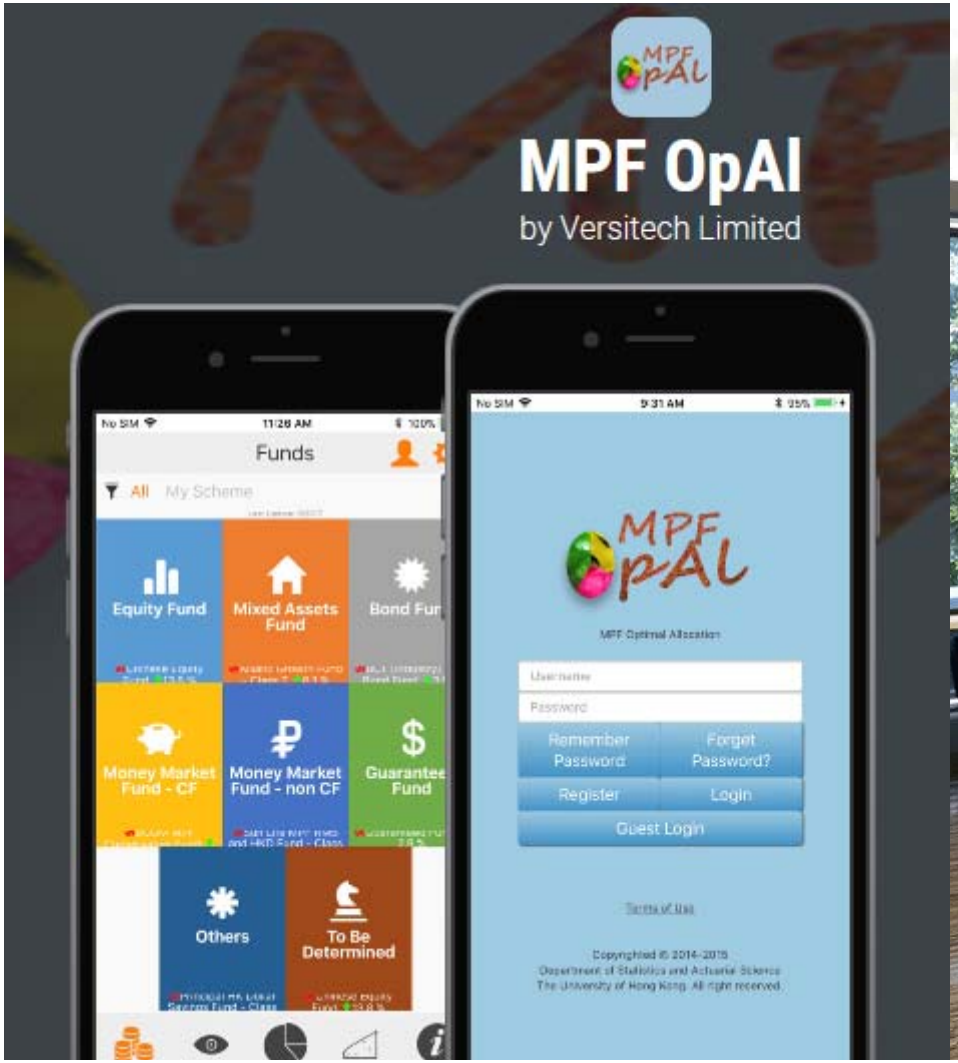
- Data Privacy and Regulatory Standards
- Security and Risk Management
- Shortage of Talents
- From Research to Deployment:

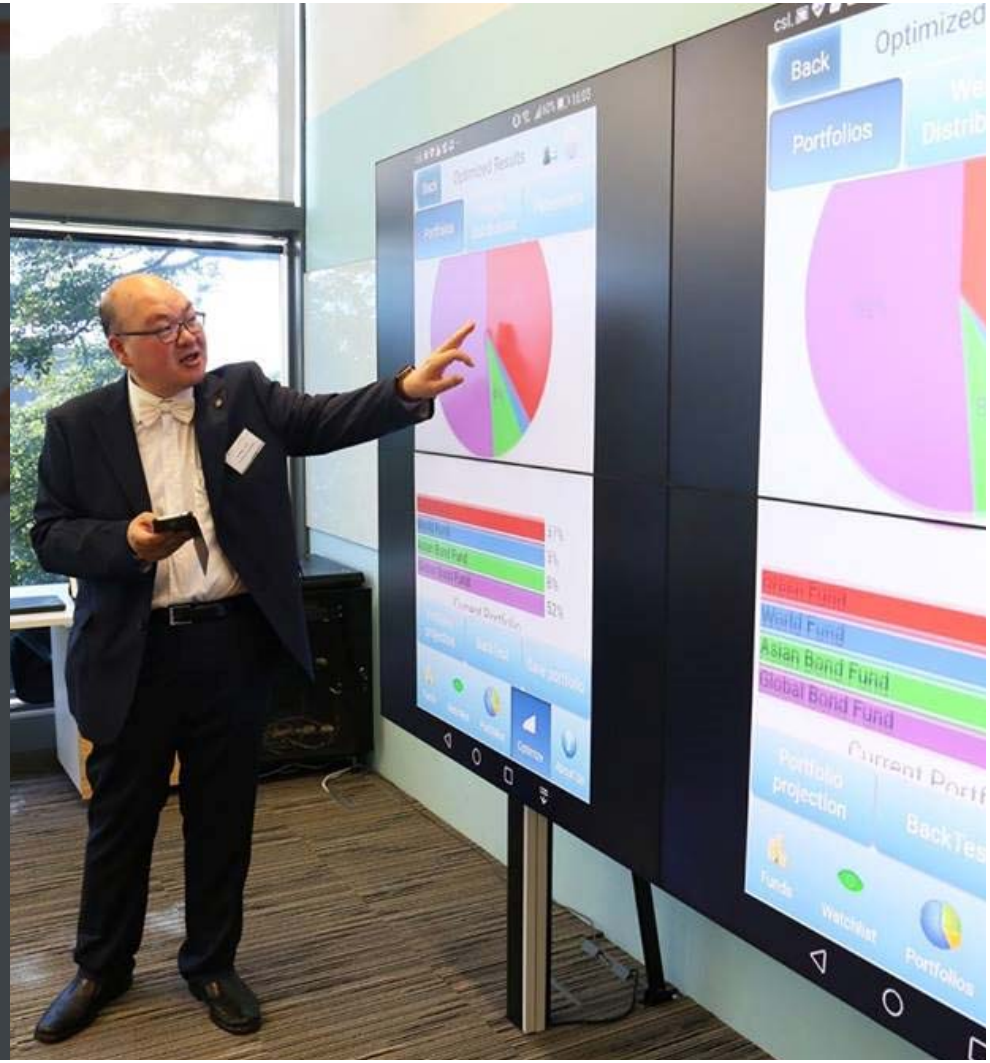
  A Long Way to Go

MPF Optimal Allocation App

Available on the App Store

Get it on Google play





Programme Overview

MORE DETAILS

The BASc(FinTech) degree is designed to nurture financial technologists and entrepreneurs with essential knowledge in both finance and technology, so they can take leading roles in innovation and applications of financial technology.

AppliedAI

BASc | HKU
Bachelor of Arts & Sciences

WHY AI? PROGRAMME INFORMATION E-BROCHURE ADMISSION OTHER INFORMATION CONTACT

BASc
Bachelor of
Arts & Sciences

香 港 大 學
THE UNIVERSITY OF HONG KONG

AppliedAI


thank you
dankegracias
merci
grazie
terima kasih
楊良河 Philip Yu
plhyu@hku.hk