摘要

近年來，隨著金融市場變動加劇，傳統投資組合理論（如馬可維茲模型）所倚賴之理性投資人與資產報酬呈常態分配等假設，在實務中面臨諸多挑戰。過去研究指出，馬可維茲模型屬於靜態結構，難以納入景氣循環等總經指標，限制其動態調整能力。為彌補此一不足，具備行動回饋機制與環境互動學習特性的深度強化學習逐漸成為金融領域研究的新興方向。此一特性使得 DRL 具備與環境互動與回饋學習能力，能整合總經指數並辨識景氣變化動態調整投資策略，展現更高的適應性與靈活性。

因此，本研究選用兩種具代表性的深度強化學習模型，分別為近端策略優化（Proximal Policy Optimization, PPO）與雙重深度 Q 網路（Double Deep Q-Network, DDQN），作為 on-policy 與 off-policy 策略的代表，並與傳統馬可維茲投資組合理論進行比較分析。再觀察資料中加入 CPI、失業率等六項總體經濟指標，並以 CLI 定義景氣循環四階段，比較三種模型在不同景氣循環階段下對 SPY、SHY、DBC 投資組合的配置行為與績效表現，檢驗其在風險調整後報酬率、策略穩定性與動態調整能力等層面的應用潛力。實證結果顯示，PPO 與 DDQN 均能根據市場與總體經濟環境的變化主動調整資產比重，展現出高於靜態配置方法的適應性與靈活性。特別是 PPO 在納入總經變數後，對景氣訊號的反應更為敏銳，能迅速調整配置方向；更重要的是， DDQN 雖未直接納入總經資料，仍可藉由市場價格學習景氣週期的變化，且其報酬率與夏普比率表現皆優於 PPO，整體而言，深度強化學習模型於策略彈性與適應能力方面，均明顯優於傳統的馬可維茲方法。

關鍵字：景氣循環、資產配置、現代投資組合理論、機器學習、深度強化學習

ABSTRACT

In recent years, financial market volatility has challenged traditional portfolio theories like the Markowitz model, which rely on rational investor behavior and normally distributed returns. Research has highlighted these limitations. The static structure of the Markowitz model also restricts its ability to incorporate macroeconomic indicators, such as business cycles. To address these challenges, Deep Reinforcement Learning (DRL) has emerged as a key area of research in finance, offering feedback and interaction with the environment. DRL allows for the integration of macroeconomic data, recognition of business cycle shifts, and dynamic adjustment of investment strategies.

This study employs two DRL models—Proximal Policy Optimization (PPO) and Double Deep Q-Network (DDQN)—to establish benchmarks for on-policy and off-policy strategies, respectively, and evaluates them against traditional Markowitz portfolio theory. The analysis includes six macroeconomic indicators, such as the Consumer Price Index (CPI) and the unemployment rate, using the Composite Leading Indicator (CLI) to delineate the four stages of the business cycle. The objective is to compare the asset allocation behavior and performance of the three models—PPO, DDQN, and Markowitz—across different phases of the business cycle for portfolios consisting of SPY, SHY, and DBC. The study assesses performance based on risk-adjusted returns, strategy stability, and dynamic adjustment capabilities.

Empirical results indicate that both PPO and DDQN can adjust asset weights in response to market and macroeconomic conditions, demonstrating superior adaptability compared to static allocation methods. Notably, PPO shows increased sensitivity to economic signals when macroeconomic variables are incorporated, allowing for rapid allocation adjustments. While DDQN does not directly integrate macroeconomic data, it learns from market price signals to identify changes in the business cycle, outperforming PPO in returns and the Sharpe ratio. Overall, DRL models exhibit significantly enhanced flexibility and adaptability compared to the traditional Markowitz approach.

Keyword: Business Cycle, Asset Allocation, Modern Portfolio Theory, Machine Learning, Deep Reinforcement Learning

目錄

表目錄

圖目錄

第一章 緒論

第一節 研究動機與問題

傳統投資策略，如馬可維茲 Markowitz (1952) 所提出的現代投資組合理論 (Modern Portfolio Theory, MPT)，建立於一系列理想化假設之上，包括投資人理性行為、以及資產報酬服從常態分配等。正如 Wang (2024) 所指出，MPT 假設市場為完全有效，市場價格能夠完全反映所有可得資訊，因此投資者能準確掌握市場風險與報酬變化；所有投資人皆為風險趨避者，且必須在同一投資期間內做出決策。然而，這些模型所依賴的假設在真實市場環境中往往難以完全成立，進而限制其在實務上的應用。儘管已有多種應對方式，如 Jin et al.(2016) 提出可透過數學程序求解器(CPLEX)等最佳化工具納入更多現實約束條件已提升模型彈性，但此類方法仍依賴嚴格的模型結構，仍受許多限制影響。而 Wang (2024) 以馬可維茲模型為基礎，納入五項現實投資限制條件，透過限制最佳化 (constrained optimization) 模型優化投資組合配置，結果顯示風險與報酬率間可達最佳平衡，強化此模型在實務操作中的可行性。作者同時指出，未來可藉由引入進階機器學習技術、擴增資產種類，並採用動態資產配置方法，以提升投資模型的適應性與實務價值。而 Chen & Zhang (2021) 提出基於機器學習與多模型架構的投資組合最適化方法，透過資產報酬預測與資產選擇，有效提升整體投資組合績效。此類方法相較於傳統依賴嚴格假設的模型驅動策略，更能因應市場情緒、政策變動等非理性因素所帶來的動態挑戰，顯示引入更具彈性且能反映實際市場動態的投資方法已成為必要趨勢。

如上所述，隨著金融市場日益複雜且快速的變動，傳統投資模型面臨諸多討戰，越來越多研究轉向數據驅動 (data-driven) 的機器學習技術 (Machine Learning, ML)，透過機器學習搭配大量廣泛的數據建構與管理投資組合，可以辨識非線性結構與高維關係，並迅速做出投資決策，協助投資人更好的進行投資組合管理 (Aithal et al. 2023)。儘管機器學習技術為金融市場帶來前瞻性的解決方案，但相對於傳統投資組合模型，仍然存在許多的爭議與討論 (Benhamou et al. 2023 & Jiang et al. 2020)。而根據 Usman (2023) 的實證研究，發現機器學習在夏普比率等風險績效指標中，表現優於傳統投資組合理論 (如馬可維茲與均分策略等)。機器學習環境中，CNN 模型回測結果穩定，雖略遜於 XGBoost，但仍優於部分傳統模型策略。該研究支持數據驅動策略作為投資組合優化的重要工具，並指出在真實市場環境下，機器學習技術具備更強的適應性與預測能力。此外，Chakravorty et al. (2018) 採用深度學習 (Deep Learning, DL) 演算法結合總體經濟資料與價量數據作為訓練數據；Obeidat et al. (2021) 則透過長短期記憶神經網路 (LSTM) 方法預測投資組合回報率。然而，這些模型都缺乏與市場的互動。換言之，深度學習在動態市場的表現也有所限制，為克服此一限制，Yang et al. (2022) 基於強化學習之演算法架構提供了解方。考量到傳統強化學習在處理高維度資料時表現有限，本研究採用深度強化學習作為主要方法論基礎。如 Yan et al.(2024) 將強化學習與深度學習相結合，應用於策略學習 (policy learning)，透過行動獎勵機制與環境互動，以處理連續時間下之投資決策問題。此類型學習的核心目標在於尋找一組能夠最大化長期回報的最

適決策序列，有效提升資產配置的經濟效益 (Durall, 2022)。相較於傳統機器學習模型強調預測之準確率，深度強化學習具備策略學習能力，能透過「行動-結果回饋」的互動過程不斷更新策略，進而在金融市場上找到最適化之資產配置。

因此，本研究採用深度強化學習的方法，結合強化學習的決策優化機制與深度學習的高維特徵提取能力，使模型得能夠在不完全資訊與延遲回饋的環境中自主探索與持續學習，此一特性適合應用於動態資產配置與即時資產管理領域 (Durall, 2022)。在深度強化學習的框架中，模型不依賴單一或特定市場機制進行學習與決策 (Buehler et al. 2019)。而 Yang (2020) 也介紹一種基於強化學習框架的集成策略以此找到最大化的投資回報率。

在此基礎上，本研究聚焦於不同強化學習策略的比較，特別是決策過程中資料使用方式對策略學習成效的影響。具體而言，強化學習可依據是否重複使用歷史資料分為 on-policy 與 off-policy 兩大類方法，各具優勢與應用條件，分別為：

1. on-policy 方法：常見的演算法包含 SARSA（State-Action-Reward-State-Action）、A3C (Asynchronous Advantage Actor-Critic)與 PPO(Proximal Policy Optimization)等，強調基於當前策略直接學習，收斂性較佳且具備策略穩定性高的特點。
2. off-policy 方法：常見的演算法包含 Q-learning、DDQN（Double Deep Q-Network）與 DDPG（Deep Deterministic Policy Gradient），利用歷史資料進行策略更新，提升樣本效率與探索能力。

根據 Yang et al. (2022) 的研究指出，在 on-policy 演算法中，更新參數後所使用的樣本無法重複利用，導致樣本效率偏低。因此， off-policy 演算法提升樣本使用效率顯得尤為重要。實證結果亦顯示，off-policy 學習在效能上優於 on-policy 方法，展現累積報酬率中具有更佳的學習表現 (Bajpai,2021)。

此外，與傳統馬可維茲最適化相比，本研究探討深度強化學習在無需預設資產分布前提下，是否能靈活融入總經指標與即時市場變動資訊的情境，並展現出更高的資產配置靈活性、風險控制能力與報酬提升潛力。基於上述脈絡，本研究希望驗證以下三點研究目的：

1. 得益於深度強化學習能靈活納入總體經濟指標數據 (Macro Variables) ，本研究預期強化學習模型能在資產配置過程中，根據經濟環境的變化自動調整資產比重，從而提升投資策略的靈活性與準確性，優於僅以歷史報酬為基礎的傳統方法。
2. 在強化學習方法的比較中，本研究將探討 on-policy 方法與 off-policy 方法在投資組合建構與風險規避策略上的適用性。考量投資決策通常涉及連續型行動空間（如資產配置權重），on-policy 方法（如 PPO）因其策略更新穩定、

樣本分布一致等特性，可能在策略優化與風險控制上表現更佳。然而，off-policy 方法（如 DDPG 或 SAC）則具備樣本使用效率高、可重複利用歷史資料的優勢，亦被廣泛應用於財務領域。為此，本研究將比較兩者在環境下的表現與穩健性。

3. 本研究亦期望證明，深度強化學習方法能夠比傳統投資組合理論在時序調整上優於傳統方法。並能在市場上做出動態資產配置調整，提升整體回報與風險控制能力，展現更高的市場適應性。

第二節 研究流程

本研究流程可分為七步驟，研究流程架構圖如圖 1 所示。首先，說明研究動機與目的，並提出深度強化學習模型在金融投資領域之應用之論點，以此探討深度強化學習方法是否優於傳統投資組合理論之資產配置。接續進行文獻回顧，探討過去研究中，傳統理論與強化學習模型於金融市場之應用成果，作為本研究模型選擇與方法設計的依據。

接續進行數據搜集與預處理，針對所有時間序列的總體經濟指標進行單根檢定，檢定是否存在單根（即是否為定態），針對非定態指標進行差分轉換後再輸入至模型中。此外，為避免產生數據洩漏 (Data Leakage)，即模型在訓練階段誤用到未來資訊，導致預測結果失真與過度擬合，本研究在輸入金融市場數據時，針對資產報酬率、標準差、共變異數以及總體經濟指標等資料，均延後一個月再輸入至模型當中。

再來，針對強化學習模型進行超參數設定與模型訓練，包含學習率 (learning rate, α)、折扣因子 (Discount factor, γ)、訓練回合數與神經網路架構等，以提升模型效能與穩定性。完成訓練後，透過平均報酬率、夏普比率等績效指標與檢定對三種模型進行評估，並探討在有無納入總經指標下，模型在權重配置上，是否具有顯著差異，並比較深度強化學習模型之績效表現。最後，根據研究結果進行總結，評估深度強化學習模型是否具備優於傳統策略的潛力，以及比較 on-policy 與 off-policy 兩模型的差異，並提出研究結論與未來研究建議。

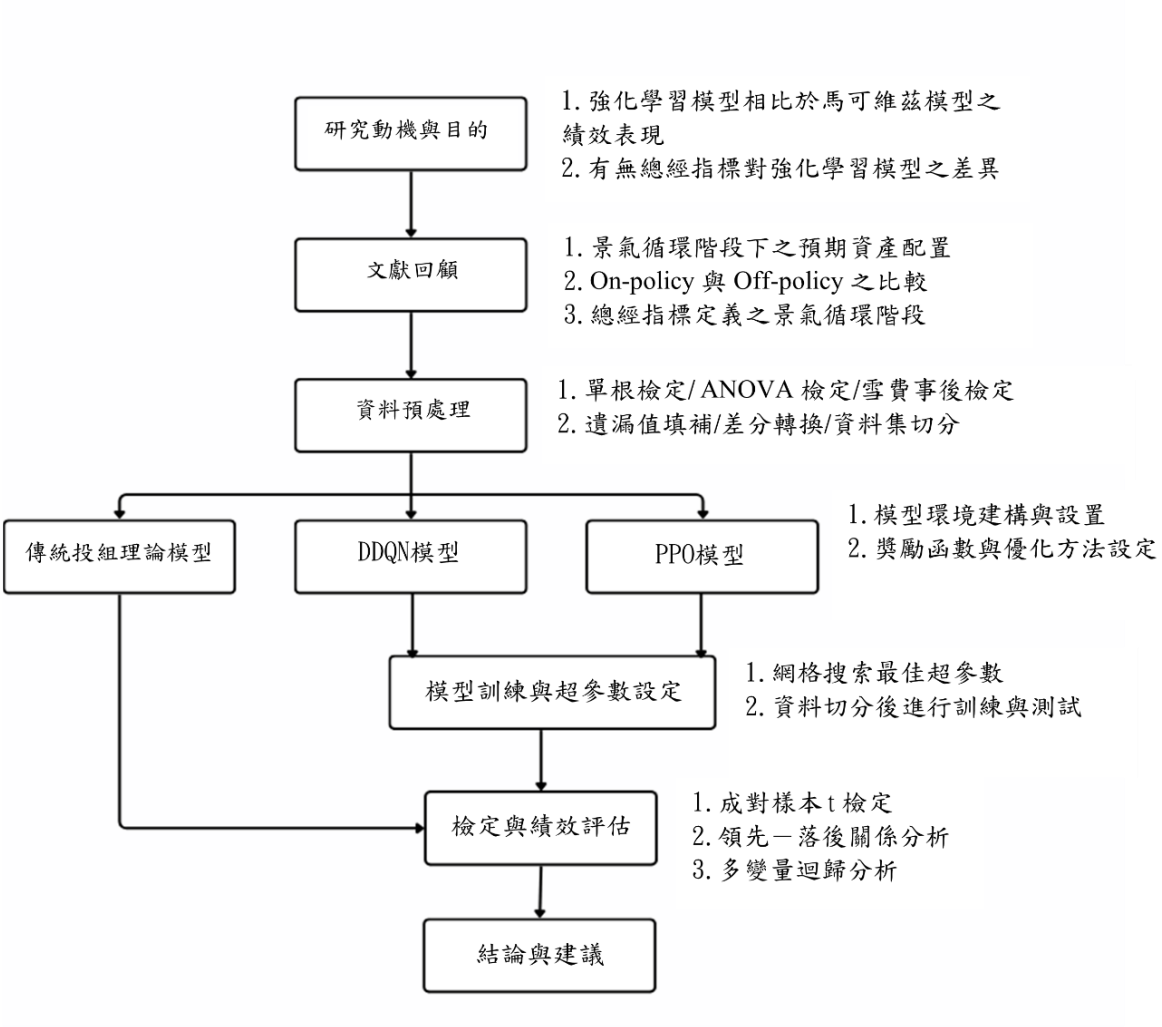
研究動機與目的
1. 強化學習模型相比於馬可維茲模型之績效表現
2. 有無總經指標對強化學習模型之差異
文獻回顧
1. 景氣循環階段下之預期資產配置
2. On-policy 與 Off-policy 之比較
3. 總經指標定義之景氣循環階段
資料預處理
1. 單根檢定/ ANOVA 檢定/雪費事後檢定
2. 遺漏值填補/差分轉換/資料集切分
傳統投組理論模型
DDQN模型
PPO模型
1. 模型環境建構與設置
2. 獎勵函數與優化方法設定
模型訓練與超參數設定
1. 網格搜索最佳超參數
2. 資料切分後進行訓練與測試
檢定與績效評估
1. 成對樣本 t 檢定
2. 領先－落後關係分析
3. 多變量迴歸分析
結論與建議

圖 1：研究流程架構圖

第二章 文獻回顧

第一節 景氣循環定義

景氣循環 (business cycles) 是總體經濟活動在時間序列上波動的過程。Burns & Mitchell (1946) 指出，景氣循環常見於以企業主導的經濟體系中，多數經濟活動會在相鄰時期內同步進行擴張 (expansion)，隨後經歷衰退 (recession)、收縮 (contraction) 與復甦 (recovery) 等四個階段。Zarnowitz (1992) 經濟週期將區分為擴張與衰退兩個主要時期，在擴張階段，整體經濟呈現正向發展，GDP 持續增長，企業營收與獲利能力提升，民間消費與私人投資意願高漲。資產價格走高，市場普遍對未來展望抱持樂觀態度。此階段常伴隨通膨壓力上升。相對而言，在收縮階段 ，GDP 成長放緩或出現負成長，消費與投資活動大幅減少，企業獲利萎縮，失業率上升，資產價格普遍下跌，政策面則可能轉向寬鬆。擴張期與收縮期將形成一個周而復始的循環，且整體週期長度可能從一年至十餘年不等。而美國國家經濟研究局 (National Bureau of Economic Research, NBER) 定義收縮期為經濟活動明顯下降，且該下降需對整體經濟造成廣泛影響並持續數個月以上。具體而言 NBER 評估景氣循環時，主要考量三大標準，深度 (depth)、廣度 (diffusion)、與持續時間 (duration)。若上述三個標準中，某一標準表現特別極端，亦可在一定程度上彌補其他面向的不足。值得注意的是，NBER 並未設定固定的指標使用規則或權重公式，而是根據實際經濟情況，並判斷各指標的適用性與重要性。

綜上所述，景氣循環的判斷方式存在多種標準與方式，反映景氣變動並非由單一變數決定，而是多項總體經濟指標共同影響。在總體經濟高度不確定的環境下，如何準確掌握景氣循環並調整資產配置策略，為投資人與資產管理者面臨的重要挑戰。總體經濟對金融市場影響深遠，資產的預期報酬率與風險結構會隨景氣週期變化而顯著變化，因此投資者需要根據不同經濟階段調整其資產配置策略。Vliet and Blitz (2011) 指出，不同資產的風險與報酬率特性受景氣變動影響劇烈，尤其在經濟由高峰轉入衰退時期，資產風險上升，對風險偏好低的投資人影響更為明顯。

為深入分析探討景氣循環對資產配置策略之影響，本文蒐整六篇具代表性的實證研究，內容涵蓋景氣階段判定方式、景氣循環階段、資產配置與研究結論加以整理與比較。此一整理有助於理解動態資產配置策略在不同經濟情境下之適用性與績效表現。下表 1 整理過去實證研究成果，以作為後續討論之基準。

表 1：景氣循環對資產配置

文獻	景氣階段劃分	景氣循環階段	資產配置	結論
Brocato & Steed (1998)	NBER 景氣	衰退期、擴張期	擴張期：提高股票類資產比重。 衰退期：提高債券與現金等固收資產。	動態調整資產配置相較靜態買進持有策略，能顯著改善報酬風險比。
Guidolin & Timmermann (2007)	資產報酬率變化	崩盤期、緩成長期、多頭期與復甦期	在崩盤期股票權重要逐漸提高。 多頭期則反向減少權重。	忽略景氣狀態的靜態配置策略顯著劣於動態配置策略。
Dzikevičius & Vetrov (2013)	CLI	復甦期、擴張期、趨緩期與衰退期	復甦期、擴張期:全額持股，甚至槓桿。 趨緩期、衰退期:減少持股，甚至退場。	股票復甦期報酬最高，衰退期報酬最低,CLI 調整可提高投資組合報酬率。
Dzikevičius & Vetrov (2012)	CLI	復甦期、擴張期、趨緩期與衰退期	復甦期：房地產、高收益股票為主。 擴張期：股票與商品比重提升。 趨緩期：黃金與商品比重提升。 衰退期:提高債券與現金比重。	使用動態資產配置模型，每一景氣階段重設期望報酬、風險與限制條件，顯著優於靜態馬可維茲模型。
Jensen & Mercer (2003)	NBER 景氣	衰退期、擴張期	擴張期：股票與房地產比重上升。 衰退期：減碼股票、配置公債與貴金屬。	衰退或緊縮時期，資產間相關性上升，分散效果變差。在擴張或寬鬆時期，資產間關聯下降，強化分散風險機會。
Lu & Su (2011).	CPI 與 GDP	蕭條期、衰退期、復甦期與擴張期	衰退期與蕭條期：債券報酬佳。 復甦期：股票為主要配置。 擴張期：期初股票仍具吸引力，末期通膨與利率調升，應轉向商品投資。	股票與商品在經濟復甦與擴張初期表現良好債券與現金於衰退末期與高利率環境中相對穩健。通膨尚未見頂前，應提前退出股票市場，轉商品或現金。

值得注意的是，傳統的靜態資產配置策略大多未能考量經濟週期變化，容易導致報酬率與風險的劇烈波動。相較之下，動態資產配置能依據不同景氣週期調整資產權重，更具彈性與穩定性。Brocato & Steed (1998) 證實，若投資組合根據景氣週期調整配置，其績效表現通常優於「買進並持有」的靜態策略，尤其在衰退期時期，投資組合中股票配置下降，債券與現金部位提高，將能有效控制風險。Guidolin & Timmermann (2007) 指出，衰退時投資組合中股票配置下降，而債券與現金部位提高，反映出投資人會主動調整投資權重以因應市場變化。Dzikevičius & Vetrov (2013) 發現，在經濟復甦期股票表現最佳，而在衰退期間則以債券表現最為穩定。此外，景氣變動不僅影響報酬率，各資產的風險水準亦有顯著差異，特別是在衰退期，所有資產報酬率標準差普遍升高，更凸顯動態資產配置的重要性。因此，在投資人能辨識所處的景氣階段的前提下，動態資產配置策略不僅能提升報酬率，更能在不同景氣時期有效管理風險，協助投資者穩健應對經濟波動。Dzikevičius & Vetrov (2012) 透過綜合領先指標 (Composite Leading Indicator, CLI) 劃分景氣為四個階段，發現股票資產表現會隨循環階段的變化而有顯著改變，並說明若於衰退期間未及時調整，將使擴張期之報酬率明顯下降。而 Koenig & Emery (1994) 指出 CLI 存在延遲性問題，特別是在景氣由擴張轉入衰退初期下降幅度有限，導致無法明確呈現轉折點，意味著無法僅使用 CLI 一個指標有效預警景氣轉折。

此外，Aruoba et al. (2011) 認為景氣循環視為無法直接觀察的潛在變數，因此使用七大工業國之月頻率總體變數建構全球共同實質經濟活動因子，能有效追蹤全球主要的經濟波動事件，並解釋跨國的經濟變異。Fossati (2016) 也認為多數景氣循環指標並未能明確預示衰退訊號，因此整合 102 個總體變數建立宏觀因子模型，其結果對於 NBER 所定義的衰退時間具有更精確的預測能力，但在擴張期間會產生假衰退信號。Chauvet & Piger (2002) 則以使用實質 GDP 成長率與就業數據建立馬可夫轉換模型，不僅能快速辨識景氣轉折點，亦能減少假陽性的情形。許多實證都顯示經濟變數在不同景氣循環階段下具有相關性 (張愷凌, 2009)，其中包含 GDP、就業、進出口貿易總值內的總體經濟指標。Kinlaw et al. (2021) 則運用製造業生產、非農就業人口等總體經濟變數之成長率，搭配相對機率與馬氏距離 (mahalanobis distance) 構建景氣循環指數，其模型對景氣變動的反應能力優於傳統 NBER 方法，並可更即時捕捉景氣階段變化。

第二節 景氣循環於資產配置之應用與限制

在實務資產配置中，面對景氣變動所帶來的市場不確定性，投資人需建構更具彈性的配置策略。因此，如何將這些景氣循環資訊有效整合進資產配置模型之中，仍為實務與研究應用的重要課題。Jensen & Mercer (2003) 指出，傳統馬可維茲模型雖能在靜態環境下達成效率前緣上的最適資產配置，但其所依賴之報酬率與風

險參數多假設為固定不變，未能納入景氣波動等總體經濟因素，限制了其於動態市場環境中的應用成效。若能將經濟週期變化納入考量，針對不同景氣階段調整資產配置權重，則有助於提升報酬潛力並有效控管風險。儘管馬可維茲模型本質上屬於靜態資產配置架構，然於實務操作中，透過滾動視窗（rolling window）方式更新預期報酬與共變異數矩陣，可使模型更貼近當前市場狀況，進而實現具彈性的動態調整效果。因此，結合馬可維茲模型之理論基礎與景氣週期資訊，不僅可強化資產配置策略之應變能力，亦有望在長期投資中提升風險調整後之績效表現。

值得注意的是，即便透過滾動視窗的方式判斷不同景氣週期的權重，單純依賴歷史資料所估計的馬可維茲最適組合，並不必然為未來的「最佳」投資組合。過去觀察到的資產期望報酬與風險，未必能反映未來的資產報酬與風險。由於金融市場具備高度隨機性，估計參數存在顯著樣本誤差，Michaud (1989) 認為均值-變異數模型最佳化方法有顯著缺陷：

1. 估計誤差極大化導致報酬率微幅變動皆會對決策造成重大的影響。

2. 某些最適組合包含極端權重，難以被投資人接受。

3. 歷史期望報酬率並非未來期望報酬率，無法完全作為預測之參考數據。

DeMiguel et al. (2009) 均值-變異數模型在實務操作時，需要估計「預期報酬率」與「共變異數矩陣」，然而這些估計值往往與實際投資期間的真實表現存在顯著誤差，導致過度依賴歷史資料所推估出的配置方法，在現實市場中難以實現理論最適，也可能因估計誤差而劣於簡單的均分策略。Gennotte (1986) 也提出預期報酬等重要變數必須透過觀察過去歷史資料進行估計，因此不可避免的產生估計誤差，並證明即使有無限多的歷史資料，估計誤差亦不會消失。因此均值-變異數模型在實務存在根本上的限制。風險趨避投資人會因估計不確定性而降低對高誤差資產的投資權重。

第三節 機器學習方法於資產配置之應用

從 2000 年起，電腦計算能力提升與金融市場數據快速增加，為了更好的解決馬可維茲模型在實務上的限制，機器學習逐漸被廣泛應用於金融市場 (Periklis Gogas, 2021)，相較於傳統財務理論，機器學習不依賴於事先設定資產報酬率的分布型態且具多輸入方法 (Multi-input method)，可納入額外的資訊來協助判斷 (Abhishek et al. 2020)。例如，Bartram et al.(2021) 將機器學習應用於主動式投資組合管理，其中以美國股市為主要持股標的。結果顯示，機器學習能有效處理股價、總體經濟等金融資料，不僅可應用於風險評估，更能因應市場波動而即時調整

投資組合。Lei et al. (2022) 將監督式學習應用於金融市場中進行分類與預測。然而，監督式學習模型對於標籤資料有高度要求，若資料量不足或標籤品質不佳，不僅會降低預測準確性，導致模型泛化性降低。

因此其他學者，如 Sen et al. (2020) 將神經網路概念融入至機器學習，採用 LSTM 模型結合神經網路來預測未來股價變化。透過分析印度股市九個不同產業中市值前五大的股票進行股價預測與建立投資組合，計算每個組合的實際報酬率、預測報酬率與風險。結果顯示，LSTM 在股價預測方面具有高度準確性，並展現更加優異的資產配置能力。而深度強化學習技術 (Deep Reinforcement Learning, DRL)，透過代理人 (agent) 在環境 (environment) 中的互動，以及動作 (action) 所帶來的獎勵進行學習。深度強化學習能持續控制與延遲獎勵 (只在整個投資周期結束時計算獎勵) 的動態決策，有效捕捉時間序列資料間的相關性，並直接面對市場條件產生對應的動作，有效適應市場環境的變化。而在深度強化學習中，演算法依據資料收集與策略學習的一致性，可分為同策略演算法 (on-policy) 與異策略演算法 (off-policy) 兩類，以下將介紹兩演算法之特色。

2.3.1 同策略演算法

On-policy 方法要求代理人僅能利用當前策略所生成的互動資料進行訓練。代表性演算法包括 TRPO (Schulman et al. 2016)、PPO (Schulman et al. 2017) 與 A3C (Mnih et al. 2016)，這些方法皆以策略梯度 (policy gradient) 為基礎，透過最大化優勢函數 (advantage function) 與最小化策略損失函數來更新 Bhardwaj 策略參數。on-policy 方法通常具備較高的收斂穩定性與理論可證性 (Zhang & Ross, 2021)，但由於每次策略更新後必須重新與環境互動以收集資料，造成其樣本效率相對較低。此特性在高維或複雜任務中尤其明顯，往往需依賴大量樣本與分散式運算資源，導致訓練成本大幅提高 (Haarnoja et al. 2018)。

在各類 on-policy 方法中， PPO 以其穩定性與學習效率的良好平衡而廣受應用，其核心設計在於限制剪裁策略更新出現大幅度更動 (Schulman et al. 2017)，此特性對於金融市場這種充滿波動與不確定性的環境來說尤其重要。透過限制剪裁策略 (Clipped Surrogate Objective) 更新，PPO 能夠避免策略因短期獎勵的波動而產生劇烈變動，在建構投資組合時，有助於維持資產配置的連續性與穩定性 (Liang et al. 2018)，避免因策略更新過於劇烈而頻繁調整持倉部位，造成交易成本上升的風險。Smith & Brown (2011) 研究指出，頻繁調整資產配置將使投資組合大幅更動，導致交易成本過高，造成風險上升之情形。因此主張透過簡化交易頻率，將可有效平衡報酬與交易成本，其核心概念與 PPO 限制策略改變的設計相互呼應。剪裁策略的設計，使得策略更新更加穩健，避免進行投資組合的過度調整。Liang et al.

(2018)、Baek (2024) 分別將 PPO 運用於中國與美國股票市場，研究顯示，PPO 在充滿雜訊與不確定性的金融市場中，不僅能學習市場的變化趨勢，還能根據不同的狀況持續調整資產配置，與傳統的馬可維茲投資組合理論相比，在報酬率與風險之間展現出更有彈性的平衡，顯示出它在金融市場與建構投資組合的潛力。

表 2 彙整了五篇應用於金融市場的深度強化學習研究文獻，聚焦於 on-policy 類型演算法在投資組合優化領域的實務應用。這些演算法基於策略梯度方法，藉由與市場環境互動逐步更新策略，學習出在風險與報酬間取得最適平衡的資產配置方式。從表 2 中可以觀察到，常見的研究標的為 S&P 500 指數 或其成分股作為代表性資產。在狀態空間設計上，on-policy 模型普遍採用與市場相關的連續型財務特徵，如報酬率、資產價格變化率與成交量等。這些變數能有效提供市場當前風險與趨勢的資訊，協助策略辨識可行的配置方向。而行為動作空間方面，則大多建構為連續資產配置比例向量，使模型能夠輸出各資產的權重分配結果，並進行動態再平衡；在獎勵函數的設計上，常見目標包含最大化夏普比率、報酬率或風險調整後績效。特別是 PPO 模型中，Sood et al. (2023)引入夏普比率變化率作為獎勵函數，以即時引導策略朝向風險調整報酬提升的方向學習。整體而言，on-policy 演算法在金融領域的應用強調即時學習與策略穩健性，透過精細的狀態空間設計與連續的動作控間配合剪裁機制，使得 PPO 能在高波動與高不確定性的市場環境中，維持穩定的資產配置邏輯與學習策略。

表 2：on-policy 深度強化學習研究文獻

文獻	演算法	標的	狀態空間	行為空間	獎勵函數	結論
Kim et al. (2019)	Asynchronous Advantage Actor-Critic (A3C)	加密貨幣資產[1]	價格變動向量	連續資產配置比例	報酬率最大化	A3C 模型在多組實驗中均優於傳統 DPG 模型。在牛市中獲得高達 17.3 倍報酬、在熊市也有 1.8 倍報酬，證實其在不同市場狀況下皆具有優秀表現。
Baek (2024)	馬可維茲 and PPO	S&P 500	投資組合總價值、商品權重	連續資產配置比例	報酬率最大化	模型在累積報酬與風險調整後指標上皆優於傳統與單一強化學習模型，顯示具備良好應用潛力，亦指出未來可提升泛化能力。
楊晴穎（2021）	A2C, PPO and SAC	iShares MSCI 美國 ESG 精選指數 ETF 成分股	商品之開盤價、最高價、最低價與調整收盤價	連續資產配置比例	報酬率最大化	模型表現優於消極式投資策略。持有 20 檔股票的情況 SAC 表現較好，持有 5 檔的情況 A2C 表現較好。
Khemlichi et al. (2023)	PPO	S&P 500	資產價格與資產報酬	連續資產配置比例	夏普比率最大化	結果顯示，PPO 在處理金融市場中高度動態與多維決策資產配置問題時，具備優異適應能力與實務應用之潛力。
Sood et al.(2023)	PPO	S&P 500	對數報酬向量	連續資產配置比例	最大化夏普比率變動量、最大回撤	研究顯示，PPO 展現出更穩定的報酬分布與較低的投資組合變動頻率，有效捕捉風險與報酬之間的平衡，進而學習出實用且穩健的資產配置策略。

[1] 加密貨幣資產如下：BTC、ETH、XRP、BCH、ETC、DASH、XMR、ZEC

2.3.2 異策略演算法

相較於 on-policy 方法，off-policy 方法則允許代理人使用由其他策略（例如過去策略或隨機策略）產生的資料進行訓練，因此在樣本重用與探索能力方面具備明顯優勢 (Sutton & Barto, 1998)。其中，Q-learning (Watkins & Dayan, 1992) 為經典 off-policy 方法，其目標為學習狀態－動作價值函數 Q(s,a)，進而導出最適策略。此類方法在資料使用效率上表現良好，但當目標策略與行為策略差異過大時，容易造成學習不穩定性與價值估計偏誤 (Clifton & Laber, 2020)。

傳統 Q-learning 演算法原本設計用於離散動作空間，在面對連續動作問題時難以直接套用，須透過額外設計可微分的策略函數以近似最佳動作。為處理此問題，部分研究採用 Actor-Critic 架構，利用策略函數 (Actor) 直接建構連續動作空間的策略，搭配價值函數 (Critic) 評估回報。然而，此類雙網路設計容易擴大策略與價值估計之間的誤差，進而加劇訓練過程的不穩定性。相對地，在離散動作空間中，DQN 結合深度學習與強化學習，為該領域帶來重要突破 (Mnih et al. 2015)，但仍承襲 Q-learning 中的「高估偏誤」 (Overestimation bias) 問題。

具體而言，由於 DQN 在最大化 Q 值時，同時使用單一網路進行動作選擇與評估，導致 Q 值容易被系統性高估 (Hasselt, 2010)。為解決此問題，Hasselt et al.(2016) 提出 DDQN，其核心創新在於將動作的選擇與評估程序分離：由主網路 (Online network) 負責在當前狀態下選擇最佳動作，反映出代理人當前的策略方向與決策偏好，再由目標網路 (Target network) 評估該動作的長期價值，從而有效降低估計偏誤，提升訓練穩定性。在應用層面， Bhardwaj et al.(2024) 將 DDQN 應用於美國股票市場，實驗結果顯示 DDQN 在累積報酬率優於傳統 DQN，展現其在高波動市場中降低 Q 值高估並穩定資產增長的能力。Bajpai (2021) 則將 DDQN 應用於印度股票市場，發現其在長期投資情境中具備較佳的風險報酬特性，尤其在市場震盪期間展現出更高的資產穩定性與策略韌性。

表 3 彙整六篇應用於金融市場的深度強化學習研究文獻，聚焦於 off-policy 類型演算法在資產配置與投資決策問題中的應用。這些研究多以 Q-learning 系列（DQN、DDQN、Dueling DDQN）與 Actor-Critic 架構下的演算法（DDPG、TD3、SAC）為核心，並實證其在不同金融標的（股票、ETF、外匯）下的效能表現。

在狀態空間設計方面，多數研究結合歷史價格數據（HOLCV）與技術指標 (移動平均線等)，有些更納入持股狀態、市場宏觀經濟變數。此設計反映出金融市場具高度時序性與多變性的特性，因此以多維資訊構建豐富的觀察空間，進而提高策略的適應性與實用性。在行為空間的設計上，部分研究使用離散型資產配置動作

（買進、賣出、持有），這不僅簡化了動作空間的設計與訓練難度，也較貼近真實交易情境中的決策方式。而在處理連續型配置比例的研究中（DDPG、TD3、SAC），其能藉由 Actor-Critic 架構有效處理高維連續動作輸出，則更符合投資組合資產配置需求，儘管設計與訓練相對複雜。

在 off-policy 眾多演算法之中，最終本研究選擇採用 DDQN 作為強化學習模型。根據 Bajpai（2021）與 Bhardwaj et al.（2024）將 DDQN 應用於金融市場的研究結果顯示，DDQN 將動作選擇與價值評估程序分離，能有效降低傳統 Q-learning 中常見的 Q 值高估問題，進而提升學習過程的穩定性與決策的可靠性。此外，相較於需同時訓練策略函數 (Actor) 與價值函數 (Critic) 的 Actor-Critic 架構，DDQN 雖同樣依賴主網路與目標網路兩個 Q 值網路（Q-networks），但其訓練流程較為簡潔，無需額外設計與優化 Actor 網路，有助於降低實作與超參數調整的複雜度，進一步提升訓練效率與模型穩定性，特別適合用於探索性研究階段及實務應用的初始部署。上述研究亦指出，DDQN 在累積報酬率、最大回撤控制等關鍵績效指標上表現良好，且在市場震盪或下行期間，能維持較高的資產穩定性與策略韌性，展現出穩定的學習性能、優異的風險控制能力與可觀的獲利潛力，凸顯其於真實交易場景中的實用價值。

表 3：off-policy 深度強化學習研究文獻

文獻	演算法	標的	狀態空間	行為空間	獎勵函數	結論
柯元富 (2022)	DDQN	台灣 0050	HOLCV、技術指標[2]	離散資產配置	交易收益	有完整資料的 0050 各股測試結果略好於買進持有策略。0050 中電子股測試結果優於買進持有策略。
黃冠棋 (2021)	DDQN	台灣 0050	HOLCV、當前持有股數	離散資產配置	交易收益最後清空持有股數	DDQN 優於基本獎勵率、CNN 模式比全連接網路模式具有更穩定。適用於研究下跌趨勢。
黃牧天 (2021)	DDPG, TD3 and SAC	美國 7 檔 ETF	利率、經濟指標	連續資產配置	財富增長率最大化	模型在投組合管理方面表現優於買進持有策略和 MVO 模型。且 DDPG 適合高報酬的客群，TD3 和 SAC 適合穩定回報客群。
Bajpai (2021)	DQN, DDQN and Dueling DDQN	印度股市[3]	HOLCV、Adj Close	離散資產配置	報酬率最大化	模型表現優於傳統方法。Dueling DDQN 整體優於 DDQN、DQN， Double DQN 優於 DQN。
Huang (2018)	DQRN	12 組外匯資產	HOLCV	離散資產配置	報酬率最大化	DRQN 顯著優於隨機策略與傳統的買進持有策略。在多數貨幣對下皆能實現正報酬，且具備較高的夏普比率與穩定的獲利表現。
Bhardwaj et al.(2024)	DQN, DDQN	AAPL, GOOGL, NEOG, K, T	HOLCV、技術指標[4]	離散資產配置	各股設定固有獎勵	DDQN 明顯優於 DQN，均優於傳統的買進持有策略。模型最小平均最大回撤（MDD）較低，表示下行風險控制良好。

[2] 技術指標為均價、均價差、交易量、交易量差、RSV、KD 指標、RSI、布林通道
[3] 印度股市資料為 TCS, RELIANCE, ZEEL, TATAMOTORS, TECHM, UPL, ULTRACEMCO, TATASTEEL, NESTLEIND, POWERGRID
[4] 技術指標為 MACD、RSI、ADX

2.3.3 同策略演算法與異策略演算法之比較

本文 2.3.1 與 2.3.2 小節分別說明 on-policy 與 off-policy 應用於金融市場的文章，下表 4 彙整七篇 on-policy 與 off-policy 兩者應用於金融市場的文獻，涵蓋演算法類型、資產標的、狀態空間設計、行為空間結構、獎勵函數與實驗結論，提供系統化比較與整理。此表不僅有助於讀者掌握不同演算法在金融應用上的設計趨勢與差異，也可作為後續模型選擇與實驗規劃的重要參考依據。在資產配置相關的強化學習研究中，on-policy 與 off-policy 方法則各自展現出不同的學習特性與應用優勢。

On-policy 方法強調代理人使用與學習相同的策略與環境互動，即從「當前策略」直接學習。這類方法具備對市場變化的即時反應能力，能快速根據環境調整行為策略，展現出良好的穩健性與泛化能力。例如，Khare et al.（2023）發現 SARSA 在測試階段比 Q-Learning 更具穩定性，而 Corazza & Sangalli（2015）亦指出 SARSA 能在高度變動的市場中迅速吸收新資訊。然而，on-policy 方法在探索空間上較為受限，導致其在策略學習效率與長期報酬表現方面可能略遜於 off-policy 方法。相對地，off-policy 方法則允許代理人從與當前策略不同的經驗中學習，例如利用歷史資料或他人策略生成的資料進行訓練，進而提升學習效率與靈活性。這種特性使 off-policy 方法在多數實驗中展現出更高的累積報酬表現。舉例來說，Pendharkar & Cusatis（2018）的研究顯示 TD Control 在長期投資報酬上優於 SARSA，而 Zhang et al.（2020）也發現 DQN 在多種金融資產，尤其是商品與外匯市場中表現出色。不過，off-policy 方法相對也對市場環境變化較為敏感，例如 Q-Learning 在市場條件變動下表現波動較大，顯示其應用上需搭配更嚴謹的風險控管與更具代表性的訓練資料。

綜合來看，on-policy 方法較適合於高度動態或不穩定的市場環境，能提供穩定且具適應力的策略；而 off-policy 方法則適用於可重複利用大量歷史資料、對長期報酬最大化有高度需求的情境，特別是在穩定性較高或資料豐富的市場中發揮出色。兩者在金融領域的應用並非彼此排斥，而應根據具體的市場條件、資料品質及投資目標加以選擇與融合。進一步分析具代表性的演算法，PPO 作為廣泛應用於金融投資場景的 on-policy 方法，其核心特點在於採用剪裁機制（clipping）限制策略更新幅度，有效避免策略崩潰，並提升學習過程的穩定性。同時，PPO 在保持策略穩健性的同時也展現出良好的樣本效率，使其特別適合應對高波動與非穩態的市場環境，展現出強大的適應性與決策靈活性。而在 off-policy 方法中，DQN 雖為強化學習的重要突破，卻存在 Q 值高估的偏誤問題。為解決此限制，DDQN 將動作選擇與價值評估過程分離，有效降低估計偏差，進一步提升訓練穩定性與預

測準確性。在金融市場應用中，DDQN 較傳統靜態投資策略更具彈性，特別在資產配置與風險控制方面展現出優異的動態調整能力。值得注意的是，隨著金融市場資料的日益龐大與複雜，強化學習演算法在資產配置中的應用也逐漸朝向多維資訊融合與動態調整發展。不論是透過技術指標、成交量、波動率，或是宏觀經濟變數等作為狀態空間的設計依據，都顯示出強化學習模型在資料處理與策略學習上的高度靈活性。這種資料驅動的特性，亦是其相較於傳統投資理論的重要優勢之一。基於上述演算法特性與應用成效，本研究選擇代表性的 on-policy 演算法 PPO 與 off-policy 演算法 DDQN 作為主要研究模型，探討其於投資組合配置與風險管理任務中的實務表現，並進一步與傳統投資組合理論進行系統性比較與分析，以驗證其在現實金融市場中的可行性與潛在優勢。

表 4：深度強化學習綜合研究文獻

文獻	演算法	標的	狀態空間	行為空間	獎勵函數	結論
Pendharkar & Cusatis (2018)	SARSA and TD Control	S&P 500, AGG 與美國 10 年期國債	discrete state	SARSA: 離散資產配置	夏普比率最大化	TD Control 代理人在多數測試期間的累積報酬率皆優異，SARSA，為強化學習在金融領域的應用提供了有力實證。
Baek (2024)	馬可維茲 and PPO	S&P 500	投資組合總價值、商品權重	TD Control 連續資產配置	報酬率最大化	模型在累積報酬與風險調整後指標上皆優於傳統與單一強化學習模型，顯示具備良好應用潛力，亦指出未來可提升泛化能力。
Liang et al. (2018)	DDPG, PPO and Adversarial PG	中國股市	HOLCV	A2C : 連續資產配置	報酬率最大化	PG 在多項實驗中優於 PPO 與 DDPG，並有效提升報酬率，顯示深度強化學習具備資產配置潛力，然而能需針對資料品質、目標函數與策略表現波動進行優化。
Corazza and Sangalli (2015)	Q-Learning and SARSA	義大利股市	近四期資產報酬率	離散資產配置	夏普比率最大化	結果顯示 Q-Learning 與 SARSA 皆具學習與獲利潛力。SARSA 對新資訊反應較快，而 Q-Learning 在高探索率下表現更佳。
Hieu (2020)	DDPG, GDPG, and PPO	AAPL, PG, BSAC, XOM	收盤價與最高價	連續資產配置	資產成長率最大化	本研究比較多種強化學習方法在股票投資組合管理上的表現，結果顯示 DDPG 在回測中表現最佳，具備穩定的收益潛力。
Khare et al. (2023)	SARSA and Q-learning	S&P 500	HOLCV	離散資產配置	報酬率最大化	SARSA 在測試階段更強的穩健性與泛化能力，Q-Learning 在訓練階段學出較優策略，但易受市場條件改變影響。此外，模型的成功高度依賴訓練資料的代表性與市場的可預測性。
Yang et al. (2020)	DQN,DDQN, DDPG, PPO, A2C, and SAC	SPY,QQQ, GOOGL, AMAZ, AAPL, MSFT and S&P 500	HOLCV 帳戶餘額技術指標	離散與連續資產配置[5]	報酬率最大化、夏普比率最大化	本研究比較多種深度強化學習演算法在股票投資組合管理任務上的表現，並應用於美股市場中具代表性的資產在多資產投資組合實驗中，DDPG 與 TD3 演算法展現出整體最優表現

[5] DQN、DDQN、DDPG 為離散資產配置比例，PPO、A2C、SAC 為連續資產配置比例

第三章 研究方法

第一節 傳統投資組合理論

現代投資組合理論由馬可維茲 (1952) 所提出。投資人可透過數學方式進行均值-變異數最適化(Mean-Variance Optimization)，於不同風險容忍程度下選出期望報酬率最佳的投資組合。

MPT 建立於兩大基本假設之上，首先 MPT 假設投資人為理性且具有風險趨避的行為者（rational risk－averter)，其決策以期望效用理論（EUT）描述投資人在風險下的行為。若多個資產擁有相同預期報酬率，投資人會選擇其風險最小的資產。只有在更高報酬率的前提下，投資人才會接受更高的風險。在不同風險情況下，都能找到最大預期報酬率的投資組合。如果將這些投資組合進行連線，就是所謂的效率前緣（efficient frontier)。

MPT 的另一核心概念為投資組合多角化（diversification)。資產間若存在異質性報酬率結構，則可透過組合權重配置降低整體風險。此風險分散效應乃源自於資產間報酬率之相關性（correlation)：若兩資產報酬率呈正相關，其價格走勢相近，難以對沖彼此波動；反之，若呈負相關或低相關，則可在資產一方波動時由另一方穩定其整體組合波動性，實現風險降低。在實務應用上，若資產之間具有高度正相關，將限制分散化效益；反之，若能納入相關性低或負相關之資產，則可有效降低組合整體波動性。本研究採用皮爾森相關係數（pearson correlation coefficient）作為衡量依存關係的統計量，其數學定義如下，

$$\rho_{xy} = E\left[\left(\frac{X-u_x}{\sigma_x}\right)\left(\frac{Y-u_y}{\sigma_y}\right)\right] = \frac{\sigma_{xy}}{\sigma_x\sigma_y}, \quad (1)$$

其中，相關係數 ρ_{xy} 為以下方程式表示，X、Y 分別代表為 x、y 資產報酬率；u_x、u_y 為各自報酬率期望值；σ_x、σ_y 為各自報酬率標準差；σ_{xy} 為兩資產共變異數。理論上 $\rho \in [-1,1]$，越趨近 -1 表示兩資產報酬率越具風險對沖能力，有助於降低整體組合風險。

為進一步驗證效率前緣與風險分散效果，本研究依據各資產的期望報酬率與風險特性，推導其投資組合的效率前緣，並分析不同風險偏好下的最適資產配置策略。整體流程可區分為兩個主要步驟。首先，須建立資產的期望報酬率與資產間的共變異數矩陣，其中共變異數矩陣可以描述資產報酬率的波動性與資產兩兩之間的相關性。第二步，根據這些參數估計值，於一組約束條件下求解最適投資組合權重，以最小化整體投資組合風險。因此，本文首先計算各項資產期望報酬率與共變異數矩陣，並運用拉格朗日乘數法（method of lagrange multiplier) 求解

最適化投資組合權重。在進入模型之前，需要先進行一些投資組合的前置說明。資產報酬率常被視為服從常態分配（如式 2 所示），投資人投資於 n 個資產時。這些資產的期望報酬率（如式 3 所示）。其中 w_i 為投資組合的權重 (portfolio weights) 表示投資人在第 i 個資產上所投入財富的比例，其比例介於 0 到 1 之間，且模型不得持有空頭部位及現金（如式 4 所示）。

$$R_i \sim N(\mu, \sigma^2) \quad , \tag{2}$$

$$E(R_p) = \sum_{i}^{n} w_i \, E(R_i) \quad , \tag{3}$$

$$\begin{aligned} &0 \le w_i \le 1, \forall i = 1,2,\cdots,n \quad , \\ &\sum_{i=1}^{n} w_i = 1 \quad , \end{aligned} \tag{4}$$

在上述模型基礎下，投資人可透過各別資產的期望報酬率與風險探討不同風險條件下的資產配置策略。本研究將探討最大夏普比率投資組合 (Maximum Sharpe Ratio Portfolio, MSR)，夏普比率由投資組合期望報酬率與無風險利率構成，指每承擔每一單位的總風險，可獲得多少單位的風險溢酬。為使投資組合達到最大夏普比率，其投資組合權重為效率前緣與資本市場線 (Capital Market Line, CML) 相切之點，該組合在所有可行組合中風險調整後的最佳選擇。其數學計算方式（如式 5 所示）。

$$\begin{aligned} &Max \quad Sharpe\ Ratio = \frac{\sum_{i=1}^{n} w_i \, E(R_i) - R_f}{\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \sigma_{ij}}} = \frac{E(R_p) - R_f}{\sigma_p} \\ &subject\ to \ \ 0 \le w_i \le 1, \forall i = 1,2,\cdots,n \quad , \\ &\sum_{i=1}^{n} w_i = 1 \quad , \end{aligned} \tag{5}$$

而根據圖 2 顯示，投資組合風險與期望報酬率之關係，其中實線代表效率前緣，反映在特定風險水準下所能達到的最適報酬率，圓點標示為最大夏普比率投資組合。為簡化模型分析，並聚焦於資產風險與報酬率的關係，本研究假設無風險利率R_f為零。因此圖中資本市場線由原點出發，其斜率為最大夏普比率，並與效率前緣相切於最適投資組合。

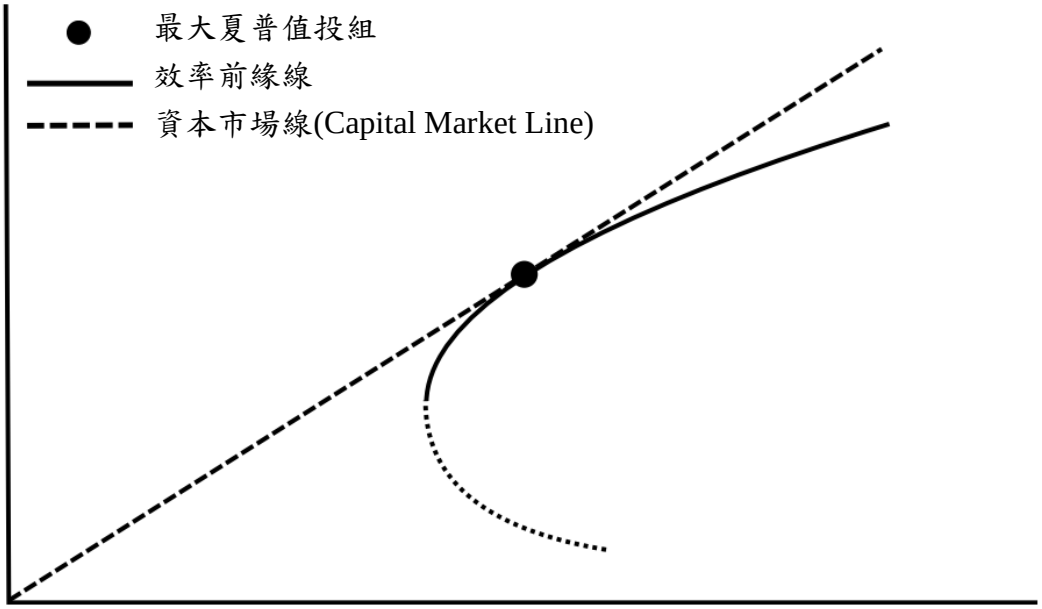

圖 2：效率前緣與最適配置示意圖

值得注意的是，為提升模型對市場變動的即時反應能力，Sengupta (1989) 在實務應用中採用滾動視窗 (rolling window) 方法動態更新預期報酬率與風險參數，反映當期市場狀況，進而強化模型在動態環境下的適應性與調適能力。此外，Fehrle (2020) 依據不同經濟循環情境調整各資產配置權重，以因應市場的非線性關係與結構性變化。

然而，儘管上述方式強化模型的即時性與彈性，但馬可維茲模型仍存在一項根本性限制，即為高度仰賴歷史資料作為預測基礎 (Guo, 2022)。實務上，資產的預期報酬率多以歷史期望平均報酬率作為替代，而風險參數則依賴歷史觀察所得的共變異數矩陣。然而，過去的市場結構與報酬率表現未必能準確預示未來。因此，以歷史平均為基礎的均值-變異數模型難以應對未來的不確定性與潛在風險的轉變。

第二節 深度強化學習理論基礎

3.2.1 深度強化學習與監督式學習之差異

Lei et al.（2022）將監督式學習應用於金融市場中進行分類與預測。然而，監督式學習模型對於標籤資料有高度要求，若資料量不足或標籤品質不佳，不僅會降低預測準確性，也可能導致模型泛化能力降低。隨著金融市場數據日益龐大且複雜，資料多為非結構性且即時變動，全面標註不僅耗費大量人力與時間，更可能無法反映市場即時變化，使得傳統監督式學習方法在實務應用上面臨挑戰。因此，深度強化學習逐漸成為處理投資組合建構的替代方案。深度強化學習並不依賴事先標籤資料，而是透過代理人與環境間的連續互動進行策略學習。代理人根據當前觀察選擇對應行動，再由環境給予回報，進一步強化或修正策略。代理人透過與環境互動、探索以及學習，逐步擬定出能最大化長期報酬的最適策略，進而應用於投資組合建構、風險控制與資產再平衡等金融決策場景。

深度強化學習核心邏輯如圖 3 所示，圍繞著代理人與環境之間的互動關係展開。整體流程可分為三個主要步驟。Sutton & Barto (1998) 將學習者和決策者稱為代理人。與其交互作用的東西包括代理人之外的一切，稱為環境。

1. 代理人會自環境中獲取一個觀察 (observation)，即環境在當下所處的狀態，可視為模型對目前情境的描述，例如：金融市場的即時數據。
2. 代理人根據這個觀察值，採取一個行動。此行動由深度神經網路所決定，目的是根據當前資訊做出最有利的決策，例如：投資組合的權重配置。
3. 行動被執行後，環境會根據該行動的結果，反饋代理人回報 (Reward)，此回報可正可負，用以評估該行動的優劣，例如：對此權重配置計算夏普比率。藉此，代理人便能透過這種「觀察—行動—回報」的反覆循環，不斷學習並優化決策策略，以最大化長期累積報酬。

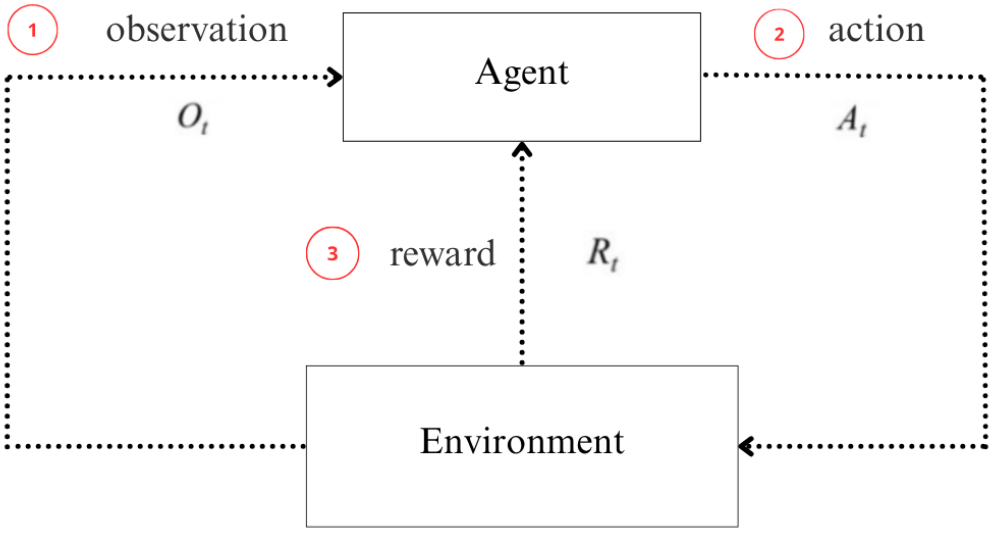

圖 3：深度強化學習架構圖

3.2.2 狀態空間

狀態空間 S 包含所有可能狀態，代理人在與環境互動時擁有的資訊。代理人在特定時間步的狀態由以下方式表示，

$$S_t = (w_i, r_t, s_t, e_t)$$

其中，t 是當前時間，w_i 為商品持有權重，r_t 是 SPY、SHY 與 DBC 之報酬率，s_t 是SPY、SHY 與 DBC 之標準差，e_t 是總體經濟指標。狀態空間用意在於彰顯影響投資決策的關鍵資訊，透過商品持有權重 (n_t)、商品的報酬與風險 (r_t, s_t) 以及消費者物價指數 (Consumer Price Index, CPI)、綜合領先經濟指標 (Composite Leading Indicator, CLI)、實質製造業與貿易銷售額 (Real Manufacturing and Trade Industries Sales, RMTS)、工業生產指數 (Industrial Production Index, IPI)、失業率 (Unemployment rate) 與初領失業救濟金人數 (Initial Claims) 六項總體經濟指標 (e_t)，使代理人能學習出具有風險控管、資產權重再平衡以及泛化能力之投資策略。詳細總經指標定義，可參考4.1節。

3.2.3 動作空間

（一） PPO 模型之動作空間設定

本研究中，PPO 模型所採用之動作空間為連續型 (continuous)。代理人於每一時間步 t 根據當前狀態輸出一組對應三檔資產的實數分數向量，再經由 Softmax 函數進行正規化轉換，以產生三檔資產對應的投資比例。此一正規化後之投資比例向量 $\mathbf{w}t = [w_1, w_2, w_3]$ 滿足以下條件：$0 \leq w_i \leq 1$ 且 $\sum_{i=1}^{3} w_i = 1$。不同於 DDQN 的離散動作設定，PPO 可於連續空間中產出任意實數型權重組合，例如 w_i =[0.183, 0.527, 0.290]，不受限於固定間距，進而提供更高的策略彈性與動態調整能力。

（二） DDQN 模型之動作空間設定

本研究採用之 DDQN 演算法，其動作空間為離散型 (discrete)，定義為代理人在每一時點 t 根據所觀察到的狀態 s_t 從一組預先定義的動作集合中選擇動作 a_t 作為資產配置策略。動作集合涵蓋 19 組離散權重選項，定義如下，

$$A = \{0.05, 0.10, 0.15, \ldots, 0.95\}$$

個別資產之配置權重從 0.05 到 0.95，以 0.05 為間距遞增或遞減，並保證整體投資組合權重總和為 1。此設計使模型能夠在維持資產配置合理性的前提下，於有限但具彈性的離散動作空間中進行策略學習與決策。

3.2.4 獎勵函數設定

獎勵函數在深度強化學習上扮演關鍵角色，不僅決定了代理人學習方向，也做為策略優化的依據。本研究採用夏普比率作為獎勵函數，使演算法在更新策略時，同時考量「報酬」與「風險」兩大核心要素。最終預期透過此獎勵函數設計，使模型能在不同市場條件下，建構出考量報酬與風險之投資組合。

第三節 深度強化學習模型：PPO 與 DDQN 的特性

3.3.1 PPO (Proximal Policy Optimization)

根據 Schulman et al. (2017) 所述， PPO 是一種基於 Actor-Critic 架構的強化學習演算法，廣泛應用於需要連續控制的任務，諸如動態調整投資組合，其核心設計在於結合 Actor 網路與 Critic 網路，以實現穩定且高效的策略學習。其中，Actor 網路負責輸出動作的機率分布，指引代理人在各種情境下做出行動決策；而 Critic 網路則估計狀態或動作的價值函數，用以評估當前策略的優劣。此外，PPO 的一大創新為引入剪裁目標函數，限制新舊策略之間的變化幅度，以避免策略更新過大所導致的不穩定性，即策略崩潰的問題。目標損失函數如下:

$$\mathbb{L}(\theta) = \mathbb{E}_t[\min(\mathrm{r_t}(\theta)\cdot \mathrm{A_t}, \mathrm{clip}(\mathrm{r_t}(\theta), 1-\epsilon, 1+\epsilon)\cdot \mathrm{A_t})], \quad (6)$$

1

其中 $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$ 用來衡量新舊策略間有多大差異；A_t是優勢函數，用衡量此動作相對於平均策略的好壞程度；ϵ 是剪裁參數，用來限制策略更新的幅度。表 5 列出完整的訓練步驟，其中第 4 列為狀態空間的設定，涵蓋資產報酬率、標準差及總體經濟指標等市場資訊；第 5 列說明動作空間的設計，PPO 採用連續權重分配方式以反映真實資產配置決策；第 6 列則定義報酬函數，以最大化夏普比率作為學習目標。演算法特點在於策略更新受到剪裁函數約束 (第 13 列)，限制更新幅度將能避免策略劇烈改變(第 14 與 15 列)，減少因過度調整投資組合而產生的交易成本與風險。

表 5：Proximal Policy Optimization（PPO）之虛擬碼

initialize: Policy parameters θ, Actor network, Critic network,
copy as old policy $\pi_{old} \leftarrow \pi$, clipping threshold ϵ
for each episode do
  for t = 0 to T-1 do
    observe current state s_t
    execute action a_t according to Actor network
    observe reward r_t
    store (s_t, $a_{t,}$ $r_{t,}$ s_{t+1})
  end for
  estimate the action value using Critic network
  compute advantage $\hat{A}_t$ using Generalized Advantage Estimation (GAE)
for epoch 1 to K do
  compute probability ratio

$$r_t(\theta) = \frac{\pi\,(a_t|s_t)}{\pi_{old}\,(a_t|s_t)}$$

  use clipping threshold ϵ and probability ratio $r_t(\theta)$ to constrain objective function

$$\mathbb{L}(\theta) = \mathbb{E}_t\left[\min\left(\mathrm{r_t}(\theta)\cdot \mathrm{A_t}\,, \mathrm{clip}(\mathrm{r_t}(\theta), 1-\epsilon, 1+\epsilon)\cdot \mathrm{A_{t,}}\right)\right]$$

  maximize the objective function to update the Actor network
  minimize the Critic loss to update the Critic network
end for
update policy $\pi_{old} \leftarrow \pi$
end for

資料來源：Firdaous khemlichi et al. (2023)

3.3.2 DDQN（Double Deep Q-Network）

在金融投資組合領域中，DDQN 適合處理離散化投資決策問題，其優勢在於能夠透過將動作選擇與評估過程分離，減少 Q 值過度高估偏誤。此特性讓演算法在動盪市場中仍能保持決策穩定性，避免因短期波動導致錯誤的投資判斷，對於設計穩健的資產再平衡策略尤其重要。

根據 Dorokhova et al.(2021) 指出 DDQN 的演算法旨在透過使用兩個獨立的估計器，分別是主網路 (Online network) 與目標網路 (Target network) 將動作選擇與動作評估分離，來解決 DQN 的 Q 值高估問題並提升學習穩定性。 Q 值高估將導致代理人更新策略時做出次優的行為選擇，從而影響收斂速度。使用主網路用來選擇在下一狀態下最有利的動作，而目標網路則負責評估該動作的價值。其目標 Q 值計算如下所示:

$$Q^{\cdot}(s_i, a_i) \approx r_i + \gamma Q_{\theta'}\left(s_{i+1}, \arg\max_{a'} \; Q_{\theta}(s_{i+1}, a')\right) \tag{7}$$

其中，$Q^{\cdot}(s_i, a_i)$ 為目標 Q 值代表預期會獲得的「實際」價值用來更新主網路的預測值，r_i 為即時獎勵 (Reward)，γ 為折扣因子 (Discount factor)。並且 $Q'\left(s_{i+1}, \arg\max_{a'} \; Q_{\theta}(s_{i+1}, a')\right)$ 是使用主網路 (Q) 選擇下一狀態下最佳動作，最後使用目標網絡 (Q') 來評估該動作的 Q 值。其中，損失函數 (Loss function) 計算如下所示，

$$L = \frac{1}{N}\sum_{i=0}^{N}\left(Q^{\cdot}(s_i, a_i) - Q_{\theta}(s_i, a_i)\right)^2 \tag{8}$$

其中，$Q^{\cdot}(s_i, a_i)$為目標 Q 值，$Q_{\theta}(s_i, a_i)$為主網路預測的 Q 值。形成預測與目標值的均方誤差，用以指導主網路的參數 θ 調整方向，透過隨機梯度下降 (Stochastic gradient descent , SGD) 進行學習。

表 6 列出完整的訓練步驟，其中，第 3 列為狀態空間的設定，涵蓋資產報酬率、標準差及總體經濟指標等市場資訊；第 5 列說明動作空間的設計，DDQN 採用離散權重分配方式以反映真實資產配置決策；第 7 列則定義報酬函數，以最大化夏普比率作為學習目標。本演算法特點在於經驗回放的引入（第 8 列），可打破樣本間的時間相關性提高數據利用效率；目標 Q 值的計算（第 13 列）採用動作選擇與動作評估分離有效降低 Q 值高估；而目標網路更新（第 16 列）則有助於穩定學習過程。

表 6：Double Deep Q-Networks Learning（DDQN）之虛擬碼

1: Initialize: Online network Q_θ and replay buffer R,
Target network $Q_{\theta'}$, with weights $\theta' \leftarrow \theta$
2: **for** each episode **do**
3: observe current state s_t
4: **for** each step in the environment **do**
5: select action $a_t \sim \pi\big(Q_\theta(s_t)\big)$ according to policy π
6: execute action a_t
7: observe next state s_{t+1} and reward $r_t = R(s_t, a_t)$
8: store (s_t, a_t, s_{t+1}, r_t) in replay buffer R
9: update current state $s_t \leftarrow s_{t+1}$
10: **end for**
11: **for** each update step **do**
12: sample minibatch N of experiences:
$e_i = (s_i, a_i, s_{i+1}, r_i)$ from replay buffer R
13: compute expected Q-values:

$$Q^{\cdot}(s_i, a_i) \approx r_i + \gamma Q_{\theta'}\left(s_{i+1}, \arg\max_{a'} Q_\theta(s_{i+1}, a')\right)$$

14: compute loss $L = \frac{1}{N}\sum_{i=0}^{N}\big(Q^{\cdot}(s_i, a_i) - Q_\theta(s_i, a_i)\big)^2$
15: perform stochastic gradient descent step on L
16: update Target network parameters: $\theta' \leftarrow \tau\theta + (1-\tau)\theta'$
17: **end for**
18: **end for**

資料來源：Dorokhova et al.(2021)

3.3.3 三模型設定差異比較

在資產配置的領域中，結合強化學習與傳統金融理論，已逐漸成為實務與學術界的重要趨勢。傳統的馬可維茲平均-變異數投資組合理論（Markowitz, 1952）雖然奠定了現代投資組合管理的基礎，但在面對複雜且高度非線性的金融市場時，其假設限制與應變能力逐漸顯現不足。相對地，深度強化學習演算法 PPO 與 DDQN，以其數據驅動的學習能力，展現出強大的潛力與靈活性。Schulman et al.（2017）提出 PPO 解決傳統策略梯度 (Policy Gradient) 方法訓練不穩定的問題，其核心設計在於限制策略更新的幅度，透過剪裁函數，使得策略調整更為平滑與可靠。這種設計的優點在於學習過程相對穩定，尤其在連續決策與風險敏感的投

資策略中，能夠有效防止策略在訓練過程中出現劇烈波動。與之對照的是 DDQN，它是一種 off-policy 的強化學習演算法，由 Hasselt et al.（2016）在解決 DQN 高估問題的基礎上提出。DDQN 能夠利用歷史經驗資料（replay buffer）進行訓練，並透過主網路與目標網路的分離架構來減少 Q 值過度估計問題，進而提升學習的穩定性。由於 off-policy 方法可以重複利用資料，其樣本效率遠高於 on-policy 方法，這對金融市場中資料成本高昂、資料量受限的情況特別重要。

相比之下，馬可維茲投資組合理論則是一種傳統的統計方法，其核心思想在於投資人基於已知的期望報酬與資產報酬率間的共變異數，透過數理規劃求解最適權重組合，在風險與報酬之間取得平衡（Markowitz, 1952）。該模型假設市場報酬服從常態分配且投資人為理性風險厭惡者，因此能夠清楚定義風險、報酬與投資偏好。然而，許多實證研究指出，金融市場中的報酬分配存在偏態與厚尾現象，違反常態假設，使得馬可維茲模型在實際操作中效果受限（Cont , 2001）。此外，當市場出現突發事件或結構性改變時，馬可維茲模型缺乏即時調整機制，需重新估算參數與求解模型，實用性大打折扣。

深度強化學習在表現上更具彈性與學習力，但其可解釋性較低，屬於「黑盒模型」，而馬可維茲模型則具備高度可解釋性與視覺化能力，能以效率前緣線清楚展現不同風險水準下的最適組合，便於與投資人進行溝通與策略說明。此外，從資料需求觀點來看，馬可維茲模型僅需輸入期望報酬與共變異數估計即可建構投資組合，而強化學習則需大量歷史資料進行反覆訓練方能穩定運行。最後，從動態調整與環境適應之能力來看，馬可維茲模型在理論教學與靜態市場環境中仍具價值，特別適合風險可量化的情境。然而，馬可維茲模型本身不具備即時更新機制，每當市場條件（如報酬率與風險）變動時，投資者需重新估算輸入參數並手動執行優化流程，難以應對高頻交易環境中快速變化的市場動態。相反的，PPO 與 DDQN 具有持續學習與即時策略更新的優勢，能根據市場變化快速反應，特別適合應用於高頻交易與多因子動態資產配置。PPO 與 DDQN 作為深度強化學習的代表，提供更強的市場適應能力與策略彈性，其中 PPO 適用於重視穩定性與風控的長期配置場景，而 DDQN 則適合運用於非線性互動複雜且資料量充足的市場結構。

下表 7 比較了馬可維茲、PPO 與 DDQN 之狀態空間、動作空間與獎勵函數，整體而言，馬可維茲模型依賴靜態統計量進行配置，強化學習模型則納入更多元的市場資訊作為輸入，並透過策略學習調整投資決策。在動作空間上，PPO 採連續型權重配置，DDQN 則為離散型策略，反映其在學習結構上的差異。三者皆以提升投資組合的風險調整報酬為核心目標，然而僅馬可維茲模型採用滾動視窗進行週期性調整，PPO 與 DDQN 則具備持續更新策略之能力，展現出其在動態市場環境中的應用優勢。

表 7：三模型設定差異比較

	馬可維茲	PPO	DDQN
狀態空間	各資產報酬率與 共變異數	各資產之報酬率、標準差、持有權重與 六項總經指標	
動作空間	連續型動作空間， 須滿足以下條件： $0 \le w_i \le 1,$ $\forall i = 1,2,\cdots,n$ $\sum_{i=1}^{n} w_i = 1$， w_i為資產配置權重	連續型動作空間， 須滿足以下條件： $0 \le w_i \le 1,$ $\forall i = 1,2,\cdots,n$ $\sum_{i=1}^{n} w_i = 1$， w_i為資產配置權重	離散型動作空間， 須滿足以下條件： $w_i \in \{0.05, 0.10, \ldots, 0.95\}$ $\sum_{i=1}^{n} w_i = 1$， w_i為資產配置權重
獎勵函數	夏普比率最大化	夏普比率最大化	夏普比率最大化
滾動視窗	使用	不使用	不使用

第四節 模型訓練流程與參數設定

3.4.1 資料切分方式

本研究將樣本期間劃分為訓練集與測試集，其中，資料之頻率維嶽資料訓練集期間為 2006 年至 2022 年，共計 196 個月，用於模型建構與參數學習，藉此讓模型能從歷史資料中學習市場規律，建立穩健的決策基礎；測試集期間則為 2022 年至 2025 年，共計 33 個月，用以評估模型在未來資料上的預測能力與實際表現，目的在於驗證模型能否有效應用於未曾見過的資料，確保其具備良好的泛化能力，並避免僅在訓練資料上表現良好的過度擬合 (Overfitting) 情形。

另外，針對馬可維茲均值-變異數模型，採用滾動視窗 (rolling window) 設計進行動態資產配置，以每 12 個月作為觀察視窗 (window size)，每 3 個月更新一次投資組合 (step size)，提升模型在市場變動下的調整彈性與穩健性。相較之下，因 PPO 與 DDQN 具備持續學習與策略更新能力，能夠根據環境即時調整決策，故本研究不對其採用滾動視窗架構。相關研究亦多採類似設計，如 Liang

et al. (2018)，並未對深度強化學習模型採用滾動視窗方式，顯示深度強化學習演算法具備持續策略調整與學習的能力。

3.4.2 強化學習超參數設定

為確保模型學習過程的穩定性與優化效果，本文採用網格搜索法 (Grid Search) 方法進行超參數組合測試。針對不同演算法選定關鍵超參數並逐一調整。表 7 整理了 PPO 與 DDQN 所設置的主要超參數與網格搜索法搜索範圍。其中，學習率決定模型參數更新的幅度，影響收斂速度與穩定性；批次大小 (batch size) 表示決定訓練所使用的樣本筆數，影響學習效率與梯度估計精度；折扣因子反映代理人對未來獎勵的重視程度。

根據 Hasselt et al.(2016) 與 McKenziea & McDonnell (2023) 描述，DDQN 架構中，ε-greedy decay 控制訓練過程中從探索 (Exploration) 轉向利用 (Exploitation) 的速度。對於 off-policy 模型而言，探索策略雖不直接影響目標策略更新，但探索樣本的多樣性對於經驗回放 (Replay memory) 中經驗品質具有關鍵作用。若探索衰減過快，可能導致樣本集中於次佳策略。在建構投資組合時過快的衰減易使模型過度配置至短期表現好的股票，降低發掘高潛力標的的能力；過慢則可能導致投資決策不穩、權重波動；經驗回放記憶體 (Replay memory size) 決定可存取的歷史交互經驗規模，對於 off-policy 模型而言，適度的記憶體容量能確保樣本多樣性，提升學習穩定性；但若容量過大，則可能包含過時策略下的資料，降低更新的時效性。在建構投資組合時，適當的經驗回放記憶體大小可保留多種市場情境下的投資報酬與行為反應資料，協助模型在不同市場環境中學習出穩健的資產配置策略，避免短視近利的投資行為。

根據 Schulman et al. (2017) 描述，PPO 架構加入了剪裁函數 (clipping parameter, ϵ) 用來限制策略更新幅度，防止在建構投資組合時，由於大幅且頻繁更動造成交易成本過高；熵係數 (entropy coefficient) 控制策略的隨機性，有助於控制模型，增加行為多樣性；廣義優勢估計 (GAE lambda) 則用於生成優勢函數估計值，提升 Critic 的穩定度與近似效果 。表 8 所列為經由網格搜索法測試後選出之最適超參數組合，其中以最大化投資組合之夏普比率作為評估指標，目的是透過系統性遍歷各組參數組合，並分別計算每個組合之報酬與標準差，進而搜索出最高夏普比例之超參數組合

表 8：模型超參數設定

超參數	敘述	網格搜索法數值	
		PPO	DDQN
學習率 learning rate (α)	調整模型參數更新速度	{0.0001, **0.0005**}	{**0.001**, 0.01, 0.1}
批次大小 batch size	每次輸入至的樣本數量	{**64**, 256}	{32, 64, **128**}
折扣因子 discount factor (γ)	模型將未來獎勵納入考量之重要程度	{**0.95**, 0.995}	{0.90, **0.95**, 0.99}
貪婪策略遞減 ε-greedy decay	控制探索率的遞減速度	-	{0.995, 0.99, **0.98**}
經驗回放記憶體大小 replay memory size	存儲代理與環境互動的記憶體大小	-	{1000, **2000**, 5000}
剪裁參數 clipping parameter (ϵ)	限制策略更新幅度	{0.1, **0.3**}	-
熵係數 entropy coefficient	控制策略探索程度	{0.001, **0.01**}	-
廣估計 GAE lambda	根據經驗生成優勢估計值，協助 Critic 逼近真實優勢	{0.95, **0.99**}	-

註: 1.PPO 無 ε-greedy decay 與 replay memory size，DDQN 無 clipping parameter (ϵ)、entropy coefficient 與 GAE lambda 2.粗體底線數字為透過網格搜索法尋找後最佳之超參數數值。

3.4.3 神經網路超參數設定

PPO 和 DDQN 模型中均為神經網路與強化學習合併應用的模型，故此類模型亦被稱為深度強化學習。加入深度學習的模型相較於強化學習之主要差異在於可以處理複雜且高維度的資料輸入與輸出(如連續型態的狀態與動作空間或大量輸入資料)，能夠更快速且精確地逼近目標函數。PPO 和 DDQN 模型分別參考了 Edmonds et al. (2018) 與 Schulman et al. (2017) 文章中神經網路超參數之設定，並整理於表9中。其中，輸入層神經元數在模型當中指的是觀測值個數，而輸出層神經元數指的是動作空間個數。由輸入層輸入觀測值後向前傳播(forward propagation) ，透過適中數量的隱藏層與其神經元進行特徵提取在輸出層產出各個動作的 Q 值，有助於提升模型一般化的能力。使用激活函數之目的在於幫助神經網路從數據中學習複雜的規律，以維持模型訓練時收斂的穩定性。方法為將數據轉換在一範圍區間(例如:線性整流函數轉換的值介於0 ~ +∞之間)，避免在更新神經元權重達到損失函數最小時，產生梯度爆炸或梯度消失

(exploding or vanishing gradient) 問題。優化器與學習率在模型訓練兩者皆會大幅影響模型收斂的速度及穩定性。前者定義更新神經網路權重的演算法，後者則影響模型參數的更新幅度，以適當的步伐逼近最佳的權重設定並滿足最小化損失函數。

表 9：模型神經網路超參數數值

超參數	敘述	預設值		
		PPO		DDQN
		Actor	Critic	
輸入層神經元數	觀測值個數	15	15	15
輸出層神經元數	動作空間個數	3	1	19
隱藏層數量	進行特徵提取	2	2	1
各隱藏層神經元數	進行提取	256, 256	400, 300	32
隱藏層激活函數	避免梯度爆炸或梯度消失	線性整流函數 (ReLU)		
損失函數	更新權重之準則	優勢函數	均方誤差	均方誤差
輸出層激活函數	避免梯度爆炸或梯度消失	線性函數	Softplus	線性函數
優化器	更新神經網路權重的演算法	Adam		
學習率	影響模型參數的更新幅度	0.005	0.005	0.007

第五節 檢定方式與績效評估

本研究比較以下模型並評估資產配置表現，分別為馬可維茲均值-變異數、PPO 和 DDQN。首先，研究認為強化學習引入總經指標進行狀態擴充與預測訓練，因此 PPO 與 DDQN 權重分配是否異於馬可維茲，表現應預期優於傳統方法。再者，研究本研究將探討 on-policy 方法與 off-policy 方法在投資組合建構與風險規避策略上的適用性，分析兩者在不同市場環境下的績效表現與穩健性。最後，本研究亦期望證明，深度強化學習方法能更早識別市場變動跡象，迅速做出動態資產配置調整，有效提升整體回報與風險控制能力，展現出相對於傳統模型更高的市場適應性。

3.5.1 總體經濟變數定態檢定

而為維持時間序列的穩定性為計量經濟模型的必要條件 Pokou et al. (2025)。因此本研究透過 ADF 檢定 (Augmented Dickey-Fuller Test) 個別檢驗不同經濟指標在時間序列中是否具有定態性，並以 10% 的顯著水準檢驗個別指標的樣本資料是否存在單根(即是否具備穩定的平均與變異數特性)，避免非定態資料導致模型學習偏誤與推論失真。若 p 值大於 0.1 則沒有足夠證據拒絕虛無假設，意味該指標存在單根，需進行差分；反之，則拒絕虛無假設。

Dickey & Fuller (1979) 的文獻中定義 ADF 檢定可以分成三個模型的統計檢定架構與其臨界值；模型一 (t_1) 為同時包含時間趨勢與截距項，模型二 (t_2) 為只有包含截距項，模型三 (t_3) 則是沒有截距項也沒有時間趨勢，並利用赤池資訊量準則 (AIC) 最小為準則選出最適落後期數（謝妁娟, 2012）。Greene (2012) 提及若錯誤使用迴歸形式將會導致檢定產生偏誤(過於嚴格或寬鬆)，造成不具單根的總經指標卻無法拒絕單根存在的虛無假設，因此 Enders (2014) 提出可先藉由視覺或趨勢迴歸分析協助判斷趨勢與截距是否存在後，選擇適合的迴歸形式進行檢定。

t_1、t_2、t_3 之迴歸形式依序呈現於下方(依序為式 9、式 10 與式 11)。式 9 為含截距項與時間趨勢項之模型；式 10 則為去除時間趨勢項之模型，僅含截距項之模型；式 11 則是不包含截距項與時間趨勢項之模型。其中 t 代表時間，α 為截距項，代表時間序列在無趨勢條件下的平均水準，β 則為時間趨勢項，衡量時間序列中是否存在趨勢，γ 則是核心檢定參數，用於檢定該時間序列資料是否存在單根(即檢定是否為定態資料)，p 是根據 AIC 最小之準則所選擇的最適略後期數。最後，ε_t 則是隨機干擾項 (假設服從獨立且同分配的隨機過程)。

$$\Delta Y_t = \alpha + \beta t + \gamma Y_{t-1} + \sum_{i=1}^{p}(\delta_i \Delta Y_{t-i}) + \varepsilon_t \quad (9)$$

$$\begin{cases} \mathrm{H_0}: \mathrm{Y_t}\text{圍繞一趨勢隨機漫步}(\gamma = 0, \beta \neq 0) \\ \mathrm{H_1}: \mathrm{Y_t}\text{為趨勢定態 (trend stationary)} (\gamma < 0, \beta \neq 0) \end{cases}$$

$$\Delta Y_t = \alpha + \gamma Y_{t-1} + \sum_{i=1}^{p}(\delta_i \Delta Y_{t-i}) + \varepsilon_t \quad (10)$$

$$\begin{cases} \mathrm{H_0}: \mathrm{Y_t}\text{圍繞一趨勢隨機漫步}(\gamma = 0, \beta \neq 0) \\ \mathrm{H_1}: \mathrm{Y_t}\text{為趨勢定態 (trend stationary)} (\gamma < 0, \beta \neq 0) \end{cases}$$

$$\Delta Y_t = \gamma Y_{t-1} + \sum_{i=1}^{p}(\delta_i \Delta Y_{t-i}) + \varepsilon_t \quad (11)$$

$$\begin{cases} \mathrm{H_0}: \mathrm{Y_t}\text{為隨機漫步}(\gamma = 0) \\ \mathrm{H_1}: \mathrm{Y_t}\text{為定態 (stationary)} (\gamma < 0) \end{cases}$$

因 t_1 之使用情境在資料具有時間趨勢的前提下進行是否具有單根的檢定，因此若無充分理由拒絕虛無假設，則認定該指標不具有單根且圍繞一趨勢隨機漫步的非定態資料。但若其中 β 之統計性並不顯著，則需去除時間趨勢項改以 t_2 進行檢定。其虛無假設設定為在截距項 (α) 具有顯著統計性的前提下，該指標圍繞於一常數隨機漫步，若有充分理由拒絕，則認定該指標為水準定態 (level stationary)。最後，若時間趨勢項與截距項均不具顯著的統計性質則需使用 t_3 進行檢定，針對該時間序列指標否存在單根之檢定進行定態與非定態資料的區別；若以上檢定之結果為非定態，則需要透過將指標轉換成一階差分值後再進行研究。

3.5.2 權重檢定方法

為了解 PPO 與 DDQN 表現是否優於傳統馬可維茲方法，本研究首先採用成對 t 樣本檢定 (paired sample t-test)，針對各模型間配置權重進行差異分析。具體而言，本文計算各模型美像資產的同期權重差異，並以 t-value 檢定資產配置是否達顯著統計水準。因此，本檢定之虛無假設為「不同模型下資產配置無明顯差異」，對立假設則為「不同模型下資產配置具有顯著差異」，若 p 值小於 0.1，則具備足夠統計依據拒絕 H_0，如下表所示，

$$\begin{cases} H_0\text{: 不同模型下資產配置無明顯差異}(w_{i,j} = w_{i,k}) \\ H_1\text{: 不同模型下資產配置具有顯著差異}(w_{i,j} \neq w_{i,k}\text{，任一等號不成立}) \end{cases}$$

其中，i 為基準模型（如 PPO），j,k 為比較模型（如 DDQN、馬可維茲），$w_{i,j}$ 為模型 i 與 j 的平均權重差異，$w_{i,k}$ 為模型 i 與 k 的平均權重差異。

表示模型間在配置策略上存在明顯差異，此檢定為後續報酬率比較以及模型反應速度（領先-落後關係）檢定建立基礎。

本研究預期三種模型在資產配置行為上應呈現顯著差異，並且反映出不同決策架構對市場資訊解讀與資產權重調整機制的異質性。

3.5.3 平均報酬率與風險績效檢定

為衡量深度強化學習相較於傳統馬可維茲方法之資產配置效果，本文採用平均報酬率與夏普比率作為主要績效評估指標。具體而言，模型組合包含「PPO + 無總經變數」、「PPO + 有總經變數」、「DDQN + 無總經變數」、「DDQN + 有總經變數」四種配置情境，並以傳統馬可維茲模型作為績效比較基準。觀察模型在長期投資下的總體表現差異（如表 10 所示）。

表 10：PPO 與 DDQN 策略結合總體變數之組合架構

	PPO	DDQN
無總經指標	PPO + 無總體變數	DDQN + 無總體變數
有總經指標	PPO + 有總體變數	DDQN + 有總體變數

透過成對樣本 t 檢定分析不同模型組合間每月平均報酬率之統計顯著性，探討其報酬穩定性與策略一致性，因此，本檢定之虛無假設為「不同模型下平均報酬率無明顯差異」，對立假設則為「不同模型下平均報酬率具有顯著差異」，若 p 值小於 0.1，則具備足夠統計依據拒絕 H_0，如下表所示：

$$\begin{cases} H_0\text{: 不同模型下平均報酬率無明顯差異 } (d_{i,j} = d_{i,k}) \\ H_1\text{: 不同模型下平均報酬率具有顯著差異}(d_{i,j} \neq d_{i,k}\text{ ，任一等號不成立}) \end{cases}$$

其中，i 為基準模型（如 PPO），j,k 為比較模型（如 DDQN、馬可維茲），$d_{i,j}$ 為模型 i 與 j 的平均報酬差異，$d_{i,k}$ 為模型 i 與 k 的平均報酬差異。

除此之外，本研究也關心夏普比率之表現，因此同樣透過成對 t 檢定分析不同模型組合間夏普比率之統計顯著性。因此，本檢定之虛無假設為「不同模型下夏普比率無明顯差異」，對立假設則為「不同模型下夏普比率具有顯著差異」若 p 值小於 0.1，則具備足夠統計依據拒絕 H_0，如下表所示，

$$\begin{cases} H_0\text{: 不同模型下夏普比率無明顯差異}\left(s_{i,j} = s_{i,k}\right) \\ H_1\text{: 不同模型下夏普比率具有顯著差異}\left(s_{i,j} \neq s_{i,k}\text{，任一等號不成立}\right) \end{cases}$$

其中，i 為基準模型（如 PPO），j,k 為比較模型（如 DDQN、馬可維茲），$s_{i,j}$ 為模型 i 與 j 的夏普比率差異，$s_{i,k}$ 為模型 i 與 k 的夏普比率差異。

本研究預期 PPO 與 DDQN 即使在無總體變數條件下績效皆能顯著優於馬可維茲，顯示深度強化學習於資產配置上的潛力。探討在不同市場條件下模型配置結果相對於基準模型的超額表現，以說明模型對於市場訊號的反應能力與策略適應性。

3.5.4 領先性分析

除了檢驗不同模型在資產配置與績效結果上的差異性外，為探討各模型配置間是否存在時間上的遞延性或反應速度，本文亦納入領先-落後關係分析 (Lead-Lag Relationship) 作為補充檢定方法，並對模型之資產權重進行相關係數進行檢定。

本文首先透過交叉相關係數 (Cross-Correlation Function, CCF) 各模型所對應資產配置變數之間的時間領先-滯後相關性，觀察在不同滯後期（如 ±5 個月）下，資產權重是否呈現顯著的正負相關。若在正滯後下出現顯著的正或負相關，表示該模型的配置結果受其他模型所領先；反之，若在負滯後 (lag < 0) 出現顯著相關，則可能表示該模型在配置反應上具有領先性，詳細公式如下：

$$\text{CCF}(k) = \frac{Cov(X_{t+k}, Y_t)}{\sqrt{Var(X_t)Var(Y_t)}} = \frac{E[(X_{t+k} - \mu_X)(Y_t - \mu_Y)]}{\sigma_X \sigma_Y} \tag{12}$$

這個公式清楚呈現 CCF 在數學上的定義，其中 $Cov(X_{t+k}, Y_t)$ 表示 X 資產在第 k 期滯後時與 Y_k 間之斜方差 (cross-covariance)，而$E[(X_{t+k} - \mu_X)(Y_t - \mu_Y)]$ 為期望值形式的結果當 CCF 在正向滯後（k > 0）下呈現顯著正相關時，代表 X 序列可能對 Y 資產具有領先性，反之亦然。

根據本研究假設，預期結果將顯示深度強化學習模型相較於傳統馬可維茲模型，以及有無總經指標之比較，應具備更佳的資訊適應能力，因此在資產配置行為上應具備明顯的差異性，進而強化其動態資產管理能力之優勢。

3.5.5 多變量迴歸分析

最後，為檢視強化學習模型在資產配置決策過程中對總體經濟變數的反應機制，本研究針對各策略於資產配置上的進行多變量迴歸分析，考量總經指標對資產配置是否具有顯著可預測之影響。考量總體經濟變數對投資行為具有滯後影響，本研究採用前期 (t-1) 總經變數預測下期 (t) 資產配置比重，以更真實地模擬決策時點與資訊可得性。此設計可用以檢視各模型對總經變數之敏感性與策略反應機制，評估其是否具有前瞻性配置能力，對模型策略與行為進行解釋。

$$w_i = \alpha_i + \beta_{i1}\Delta CPI + \beta_{i2}CLI + \beta_{i3}\Delta RMTS + \beta_{i4}\Delta IPI + \beta_{i5}UMP + \beta_{i6}Initials + \varepsilon_i\,,\ \ i = 1,2,3 \tag{13}$$

其中 w_i 分別為 SPY、SHY、DBC 之配置權重，α_i 為截距項，β_{i1} 為 ΔCPI 迴歸係數，β_{i2} 為 CLI 迴歸係數，β_{i3} 為 ΔRMTS 迴歸係數，β_{i4} 為 ΔIPI 迴歸係數，β_{i5} 為 UMP 迴歸係數，β_{i6} 為 Initials 迴歸係數，ε_i 為誤差項，而 β 的正負方向亦對應資產權重的調整邏輯，並表示資產配置對該變數具敏感性。

第四章 資料說明

本研究所使用資料分為兩大類，分別為總體經濟指標與金融資產報酬資料。其中，本研究主要依照 Dzikevičius & Vetrov (2012) 之景氣循環定義方法將景氣分為四個階段：當 CLI 大於 100 時，相較上一期呈現上升趨勢為擴張期；反之呈下降趨勢則是回落期。當 CLI 小於 100 時，相較上一期呈現上升趨勢為復甦期；反之呈下降趨勢則是衰退期。以此定義進行不同景氣循環階段下，各金融資產是否能夠有效分散系統性風險以及總體經濟指標在平均上是否具有顯著差異，並且能夠作為深度強化學習模型之狀態空間中重要的觀察，影響模型決策行為，分述如下。

第一節 總體指標資料介紹

總體經濟變數作為 DRL 模型狀態空間的一部分，目的在於輔助模型理解市場景氣變化與風險環境，透過模型自我探索與學習，期望模型可以在經濟合作暨發展組織 (OECD) 判斷的景氣循環中做出更優的判斷。表 11 整理了常用於判斷景氣變化的總體經濟指標，因此本研究從領先、同時、落後指標中，選取六個常被用於判斷景氣循環的總體指標，包含 (Consumer Price Index, CPI)、綜合領先經濟指標 (Composite Leading Indicator, CLI)、實質製造業與貿易銷售額 (Real Manufacturing and Trade Industries Sales, RMTS)、工業生產指數 (Industrial Production Index, IPI)、失業率 (Unemployment rate) 與初領失業救濟金人數 (Initial Claims)。資料來源主要源自於聯邦儲備經濟數據 (Federal Reserve Economic Data；FRED) 以及 OECD 公布之數據（可參照表 12）。本研究蒐集的資料期間為 2006 年 4 月至 2025 年 3 月，資料頻率為月資料。

表 11：總體經濟變數與景氣判斷之相關文獻

相關文獻	消費者物價指數 (CPI)	綜合領先指標 (CLI)	實質製造業與貿易銷售額 (RMTS)	工業生產指數 (IPI)	失業率 (UMP)	初領失業金人數 (Initials)
張愷凌 (2009)	-	-	-	V	V	-
Aruoba et al. (2011)	-	-	V	V	-	V
Dzikevičius & Vetrov (2012)	-	V	-	-	-	-
Fossati (2012)	V	-	V	V	V	V
Gampaoli et al. (2024)	V	V	-	-	-	-
Kinlaw et al. (2020)	-	-	-	V	-	-

表 12：總體經濟變數之資料來源與期間

總體經濟變數	資料來源	資料期間
消費者物價指數	FRED (CPIAUCSL)	2006M4-2025M3
綜合領先經濟指標	OECD (111.SLRTTO01.IXOBSA)	2006M4-2025M3
實質製造業與貿易銷售額	FRED (CMRMTSPL)	2006M4-2025M3
工產指數	FRED (INDPRO)	2006M4-2025M3
失業率	FRED (UNRATE)	2006M4-2025M3
初領失業救濟金人數	FRED (ICSA)	2006M4-2025M3

註：資料來源中 () 內為資料庫搜尋代號

由上述資料來源抓取資料後，並無遺漏值且大多數均為數值型態之時間序列資料。為維持時間序列的穩定性，必須根據選擇出來的六項總體經濟指標進行 ADF 檢定，結果呈現於表 13。檢定結果顯示，在 10% 顯著水準下，CPI、RMTS 與 IPI 之 P 值皆顯著大於 0.1，表示無法拒絕存在單根的虛無假設，意味著上述三項變數皆存在時間序列的不穩定性，故以差分的方式進行轉換並做後續分析。

表 13：總體經濟指標之 ADF 檢定

總體經濟指標	t_1	t_2	t_3
CPI	-0.21 (0.9926)	2.02 (0.9999)	4.61 (0.9999)
CLI	-3.46 (0.0468**)	-3.49 (0.0093***)	-0.10 (0.6473)
RMTS	-3.13 (0.1621)	-0.59 (0.8698)	1.06 (0.9248)
IPI	-3.00 (0.1354)	-2.44 (0.1312)	0.11 (0.7172)
UMP	-3.30 (0.0686*)	-2.67 (0.0809*)	-0.95 (0.3060)
Initials	-6.82 (<0.0001***)	-6.83 (<0.0001***)	-4.40 (<0.0001***)

註:1.落後期的選擇依據 AIC 之最小準則所設定，落後期數均為 1

2.ADF 檢定之三種模型之 ADF 統計依序為t_1、t_2、t_3，皆四捨五入至小數點後兩位。括號內為 p 值，皆四捨五入至小數點後四位

3. t_1 統計水準(臨界值): -3.13(10%),-3.43(5%),-3.99(1%)

4. t_2 統計水準(臨界值): -2.57(10%),-2.88(5%),-3.46(1%)

5. t_3 統計水準(臨界值):-1.62(10%),-1.95(5%),-2.58(1%)

6.*表示顯著水準 p <10%；** p <5%；*** p <1%

7. ΔCPI:消費者物價指數；CLI:綜合領先指標；ΔRMTS:實質製造業與貿易銷售額(百萬美元)；ΔIPI:工業生產指數；UMP:失業率；Initials:初領失業救濟金人數

表 14 則顯示六項總體經濟指標在不同景氣階段 (擴張期、回落期、復甦期與衰退期)下的敘述統計量變異情形，並透過單因子變異數分析 (ANOVA) 檢定各變數在不同景氣階段是否具有顯著差異。表 15 則透過雪費檢定，明確找出各階段間是否存在差異，以下針對各變數進行說明。

ΔCPI之平均與標準差於景氣階段間具有顯著差異，同時也符應實際情況。當景氣處於相對熱絡的階段（擴張期、回落期），總體需求上升促使商品和勞務價格上升，推動價格面之原物料價格上升，使得廠商將生產成本轉嫁於消費者，推升消費者物價指數的上升，造成高消費者物價指數成長率且遠高於相對冷卻階段(衰退期與復甦期)。於雪費事後檢定結果顯示，在景氣相對熱絡階段皆對於相對冷卻階段，例如擴張期與衰退期、復甦期與衰退期，各階段組合間皆具有統計上的顯著差異。從此指標可觀察到，在經濟相對疲軟階段 (衰退期與復甦期)，市場波動程度相對較高，反映出此期間較易受到重大金融事件、政策干預或市場情緒波動等外部因素的影響。

CLI 則涵蓋多個領先性經濟變數加權所組成，包含國際金融、勞動市場與消費者信心水準等綜合面向的資訊，為最有效輔助辨別景氣階段的總經指標。CLI 上升，預示未來幾個月經濟邁向擴張期的可能性高，而高檔持平則意味景氣狀況仍強但動能趨緩，這也導致回落期之平均略高於擴張期；反之，CLI 下降則隱含景氣轉差或衰退風險升高，於低檔的反轉回升則有望邁向復甦。作為最具有辨別景氣階段的指標，CLI 於各景氣階段的平均值具有最顯著的差異，特別是擴張期和回落期兩個相對熱絡的景氣階段都分別與衰退期和復甦期具有統計上的顯著差異，與經濟上的預期結果一致。

$\Delta RMTS$ 指的是經過通貨膨脹調整 (扣除物價變動因素) 後的實質銷售額 (百萬美元) 之差分值，反映實際交易量與景氣的強弱程度。於復甦期表現最佳（平均值為 0.0050），而在衰退期則為負值 (-0.0046) 顯然符合經濟預期。於復甦期間，需求於衰退期的谷底快速反彈回升，需求的報復性反彈致使企業上游訂單徒增，不得不重啟產能引發一段補貨潮，使得銷售額成長率明顯快速拉升。直到擴張期，雖然持續成長但比較基期已相對到達高點，導致月增率逐漸降低，於回落期更是趨於溫和成長。在雪費事後檢定也可以看到復甦期與相鄰的兩個景氣階段(回落期與衰退期)在平均上具有顯著的差異，意味該指標能夠有效辨別景氣轉折點位置。

ΔIPI 為工業生產指數，是觀察一國製造業、礦業與公營事業活動強弱的關鍵指標，也常被視為景氣同步指標之一。在復甦期與擴張期皆為正值，而於衰退期顯著下降至負值 (-0.0064)，標準差亦明顯增加，顯示該指標受景氣波動影響較大。易受到國內外需求、原物料與零組件供應情況、產能利用率等因素影響，相較$\Delta RMTS$來說，能夠更即時反映製造業與實體經濟活力。雪費檢定結果顯示，衰退期與其他的景氣階段在平均上皆具有相當顯著的差異，意味在判斷衰退期間，ΔIPI具有極高的敏感度。

在 UMP 方面，各階段的平均 UMP 之間具有統計上顯著的差異，這也符合經濟上的預期；當經濟陷入衰退期間時，勞動市場的不確定性明顯上升，反映企業在衰退期面臨需求驟降與營運壓力，不同產業發生裁員行為的規模具有差異且加上政府補助、振興政策的施行存在時間滯後，導致 UMP 於該階段平均值較高。相對而言，擴張期與回落期因經濟活動穩定回升使得 UMP 趨於緩和，平均明顯下降。透過檢定結果顯示，UMP 不僅能反映就業市場對景氣變化的敏感度，該指標更揭示勞動市場在不同經濟環境下的平均變化程度。

最後在 Initials 方面，相較於 UMP 反映就業市場之自然失業與循環性失業，Initials 更能反映在經歷不同經濟循環週期下，因景氣波動所導致首次失業的情形，為判斷景氣變化的領先指標之一，但不確定因子較高。於復甦期人數最高（473,064 人），在回落期則相對較低（300,500 人）。亦符合在經濟相對擴張階段(擴張期與回落期)，勞動市場需求增加、普遍預期市場前景看好等因素造成初

領失業救濟金人數相對較低的經濟預期；反之，在相對疲軟階段(復甦期與衰退期)亦成立，造成兩階段有明顯差異，特別是在回落期與復甦期之間。

表 14：總經指標之敘述統計量

敘述統計量 (單位)	擴張期 (n=58)	回落期 (n=46)	復甦期 (n=75)	衰退期 (n=49)	F 值 [Pr(>F)]
ΔCPI(%)	0.0022 (0.0023)	0.0036 (0.0034)	0.0019 (0.0018)	0.0007 (0.0045)	7.432*** [<0.0001]
CLI(%)	100.8524 (0.5942)	100.8770 (0.4830)	98.9985 (1.2026)	998.6659 (1.5514)	67.44*** [<0.0001]
$\Delta RMTS$(百萬美元)	0.0023 (0.0097)	-0.0007 (0.0064)	0.0050 (0.0136)	-0.0046 (0.0192)	5.794*** [0.0008]
ΔIPI(指數)	0.0024 (0.0074)	0.0001 (0.0051)	0.0033 (0.0100)	-0.0064 (0.0216)	6.869*** [0.0002]
UMP(%)	0.0567 (0.0159)	0.0497 (0.0117)	0.0660 (0.0262)	0.0574 (0.0250)	5.929*** [0.0007]
Initials(人數)	341974 (135544)	300500 (64618)	473064 (581856)	357928 (312656)	2.526* [0.0584]

註：1.F 值後之星號表示該變數於各景氣階段間具有統計顯著性差異，其中，

2.表內數值為:平均數(標準差)，僅最後一行為 F 值[p 值]

3. ***代表 p-value < 1%、**代表 p-value < 5%、*代表 p-value < 10%

4.ΔCPI:消費者物價指數差分值；CLI:綜合領先指標；ΔRMTS:實質製造業與貿易銷售額差分值(百萬美元)；ΔIPI:工業生產指數差分值；UMP:失業率；Initials:初領失業救濟金人數

表 15：總體經濟指標之事後檢定

		該期間數值顯著較高			
	景氣階段	擴張期	回落期	衰退期	復甦期
該期間數值顯著較低	擴張期				UMP
	回落期				ΔRMTS, UMP
	衰退期	ΔCPI, CLI ΔRMTS, ΔIPI	ΔCPI, CLI ΔIPI,Initials		ΔRMTS, ΔIPI
	復甦期	CLI	ΔCPI, CLI		

第二節 標的資料介紹

本研究透過分析報酬結構與資產間的共變異關係，尋求具備風險控制效果的最適資產配置。在進行資產配置優化之前，先計算各資產間的相關係數，以評估其潛在的分散化效果。資產選取則參考 Wongsawatgul (2019) 等人之研究，挑選具備組合建立潛力的標的，作為投資組合的構成基礎。

由表 16 可見，在所有樣本期間 SPY 與 SHY 之間呈現顯著負相關 (-0.2059)，SPY 與 DBC 之間亦顯著負相關 (0.3406)。此外，SHY 與 DBC 間亦呈現統計顯著之負相關(-0.9042)。整體而言，三項資產皆存在一定程度之反向變動，顯示其在投資組合配置上具備風險分散效果。所有檢定皆基於 229 個月度觀察值 (N = 229)，相關係數之顯著性以雙尾 t 檢定評估，顯著水準標記於表中。

表 16：所有時期下資產之相關係數 (N=229)

所有樣本	SPY	SHY	DBC
SPY	1	-	-
SHY	-0.2059*** (0.3523)	1	-
DBC	0.3406*** (0.2892)	-0.9042*** (0.2514)	1

註: 1.括號內數字為估計標準誤; 2.***代表 p-value < 1%、**代表 p-value < 5%、*代表 p-value < 10%

然而，僅觀察整體期間的相關係數，尚不足以全面評估資產間的關聯性。因此，為深入探討在不同景氣情境下的資產互動關係，本研究將樣本分為四種景氣循環階段，分別檢定其相關係數變化，如表 17 所示。

首先，在擴張期中，SPY 與 SHY 間呈現顯著負相關 (-0.1488)，而 SHY 與 DBC (0.2019) 間亦為顯著負相關，顯示該階段中債券與其他資產間具備穩定之分散效果。然而，SPY 與 DBC 雖呈正相關 (-0.1836)，但檢定結果不具統計顯著性，表示其間聯動性不足以視為結構性關係，該結果亦可能源自樣本波動與暫時性市場共同波動所致。

而在回落期間，SPY 與 SHY 的負相關性 (-0.1976) 加深，而 SPY 與 DBC 間則轉為顯著正相關 (0.2322)，反映市場轉折時股市與商品價格之變動方向出現背離、表現分化之現象。SHY 與 DBC 間則僅呈微弱負相關 (-0.0455) 且不具顯著性，顯示該階段二者變動關係不穩定、缺乏系統性關係。

進入衰退期，SPY 與 SHY 負相關性 (-0.1976) 最為明顯，顯示投資人風險偏好顯著下降時，債券資產扮演明確避險角色。SPY 與 DBC 間亦呈顯著正相關 (0.2322)，可能與通膨壓力導致商品價格上漲有關。SHY 與 DBC 則維持非顯著負相關 (-0.0455)，顯示其結構性，不受景氣階段轉換干擾。

在復甦期，三項資產對皆呈現高度顯著的相關性，其中 SPY 與 SHY 負相關 (-0.1981) 顯著，而 SPY 與 DBC 的正相關 (0.4519) 達全期最高，顯示市場預期回溫時，股市與商品同步上揚；同時，SHY 與 DBC 間亦呈現顯著負相關 (-0.1265)，反映出債券與實體資產表現方向一致背離，強化資產間的分散性。

綜上所述，SPY 與 SHY 於四個景氣階段皆呈穩定顯著負相關，強化其避險與分散配置角色；SPY 與 DBC 則在回落、衰退與復甦階段表現出明顯聯動關係，尤以復甦期相關性最高。DBC 則在部分期間與股票、債券無顯著關係，顯示其具備非聯動的特性，適合作為風險對沖工具。整體而言，資產間之分散效果不僅存在，且於經濟波動較大時更為顯著，驗證多元資產配置於不同總體情境下的穩健性。

表 17：所有時期下資產之相關係數

擴張期 (n = 58)	SPY	SHY	DBC
SPY	1	-	-
SHY	-0.1488*** (0.2822)	1	-
DBC	0.2019 (0.2674)	-0.1836*** (0.2013)	1
回落期 (n = 46)	**SPY**	**SHY**	**DBC**
SPY	1	-	-
SHY	-0.1976*** (0.4262)	1	-
DBC	0.2322*** (0.2890)	-0.0455 (0.2701)	1
衰退期 (n = 50)	**SPY**	**SHY**	**DBC**
SPY	1	-	-
SHY	-0.2842*** (0.3899)	1	-
DBC	0.4298*** (0.2646)	-0.1560*** (0.2744)	1
復甦期 (n = 75)	**SPY**	**SHY**	**DBC**
SPY	1	-	-
SHY	-0.1981*** (0.3207)	1	-
DBC	0.4519*** (0.2569)	-0.1265*** (0.2436)	1

註: 1.括號內數字為估計標準誤; 2.***代表 p-value < 1%、**代表 p-value < 5%、*代表 p-value < 10%

第五章 實證結果與結論

第一節 策略績效比較

為評估深度強化學習策略 (PPO、DDQN) 與傳統均值-變異數模型 ，在資產配置表現上的差異，本研究依序進行以下四項分析：首先，檢驗三種策略於不同資產的平均配置權重是否存在顯著差異；其次，比較各模型於樣本期間內的績效與風險表現，分別以平均報酬率與夏普比率為代表。並透過交叉相關分析，探討模型間在各項資產配置行為是否存在時間依賴性，最後進行多變量迴歸探討總經指標對模型權重配置的影響。

5.1.1 模型權重分配變異數分析

本研究針對三種策略資產配置結果進行比較（如表 18 所示），發現其資產配置行為展現出高度異質性，且多數差異具統計顯著性，顯示 DDQN 偏好風險資產，配置邏輯與傳統模型差異最為明顯。在 SHY 資產方面，馬可維茲配置最高，PPO 次之，DDQN 最低，反映三模型對低風險資產評估分歧，DDQN 對避險功能反應最弱。最後在 DBC 資產方面，馬可維茲偏好大宗商品，顯示強化學習模型在高波動資產上反應不一，PPO 具彈性調整能力，DDQN 則呈現迴避傾向。此外，納入總體經濟變數後（如表 19 所示），PPO 策略在 SPY 與 SHY 配置皆顯著提升，展現其對總經指標具高度感知能力；反觀 DDQN 策略整體配置變化幅度小，多數不具顯著性，顯示其學習行為主要依賴市場自身交易資訊 (如報酬率與共變異數等)，即可捕捉總經資訊。

整體而言，總經變數納入後有效強化 PPO 對景氣訊號的學習與調整行為，提升其資產配置之前瞻性；DDQN 則展現出對總經變數反應鈍化的特性，顯示不同強化學習架構於訊號感知與配置邏輯上存在根本性差異。

表 18：模型不同資產配置之權重差異檢定結果

資產	比較組合(i-j)	平均差(i-j)
SPY	PPO-馬可維茲	0.1878***
		(0.0621)
	DDQN-馬可維茲	0.6378***
		(0.0670)
	PPO-DDQN	-0.4499***
		(0.0525)
SHY	PPO-馬可維茲	-0.0767
		(0.0820)
	DDQN-馬可維茲	-0.1852**
		(0.08731)
	PPO-DDQN	0.1086***
		(0.0384)
DBC	PPO-馬可維茲	-0.1111
		(0.0751)
	DDQN-馬可維茲	-0.4525***
		(0.0787)
	PPO-DDQN	0.3414***
		(0.03173)

註: 1.括號內數字為估計標準誤; 2.***代表 p-value < 1%、**代表 p-value < 5%、*代表 p-value < 10%

表 19：PPO 與 DDQN 策略「有／無總經變數」顯著性檢定結果

資產	比較組合 (i-j)	平均差(i-j)
SPY	PPO 有總經-無總經	0.1316***
		(0.0469)
	DDQN 有總經-無總經	-0.0306
		(0.0527)
SHY	PPO 有總經-無總經	-0.0791**
		(0.0416)
	DDQN 有總經-無總經	0.0162
		(0.0494)
DBC	PPO 有總經-無總經	-0.0525
		(0.0479)
	DDQN 有總經-無總經	0.0143
		(0.0179)

註: 1.括號內數字為估計標準誤; 2.***代表 p-value < 1%、**代表 p-value < 5%、*代表 p-value < 10%

5.1.2 報酬率與夏普比率檢定結果

確認三種模型於資產配置行為上存在顯著差異後，進一步檢視其於樣本期間內之平均報酬率與夏普比率表現，以評估不同策略在整體績效上的相對優劣。為確保比較一致性，使用配對樣本 t 檢定分析三組模型對報酬率差異，其結果如表 20 所示。DDQN 策略表現最優，平均報酬率為 1.19% 顯著高於 PPO 的 0.62% 以及馬可維茲-1%，三者間差異均具統計顯著性，顯示 DDQN 具備穩定捕捉市場趨勢與報酬波動的能力。而 PPO 仍優於馬可維茲，反映其具備初步的彈性調整能力，但學習與獎勵機制尚不成熟。此外由表 21 可知，PPO 在納入總經後，夏普比率明顯提升，平均報酬由差異達顯著水準，顯示總經指標有助於強化其學習穩定性與決策準確度。同樣參照表 21，雖然 DDQN 納入總經變數後，其平均報酬率略微下降，但報酬變異幅度亦隨之縮小，使得夏普比率由 2.12 提升至 2.31（雖未達統計顯著水準），顯示其風險調整後的整體績效仍有所提升。此結果與本研究所強調之風險控管與穩健學習目標相符，反映總經變數的導入有助於提升 DDQN 模型在資產配置上的穩定性與效率，即使報酬略降，但仍能展現出良好的報酬品質與風險承擔能力。

綜上所述，DDQN 策略於樣本期間整體報酬與夏普比率最優，儘管部分指標未達統計顯著，然其在各面向之表現均衡，具體顯示其作為資產配置模型的應用潛力。PPO 策略則展現出對總經變數的高敏感度，具備調整與優化潛力，需強化模型架構、策略設計與學習架構。整體結果顯示，總經指標可作為提升策略穩定性的潛在工具，但其成效仍高度依賴模型設計與指標整合機制之相容性。

表 20：三種模型投資組合績效檢定 (n=33)

	馬可維茲	PPO	DDQN	兩兩差異比較		
績效指標	(1)	(2)	(3)	(2)-(1)	(3)-(1)	(3)-(2)
平均報酬率	-0.0001	0.0062	0.0119	0.0062 (0.0060)	0.0120 (0.0080)	0.0057 (0.0082)
夏普比率	0.3142	1.4689	2.3096	1.1556 (1.0096)	1.9954* (1.1177)	0.8406 (1.1152)
平均變異數	0.0000454	0.0000486	0.0000719	-3.15e-06 (-1.13e-06)	-2.65e-06** (-1.32e-60)	0.0000233 (-1.42e-05)

註: 1.括號內數字為估計標準誤; 2.***代表 p-value < 1%、**代表 p-value < 5%、*代表 p-value < 10%

表 21：深度強化學習模型「有／無總經變數」之投資組合績效檢定 (n=33)

	PPO			DDQN		
	無總經	有總經	差異	無總經	有總經	差異
績效指標	(1)	(2)	(2)-(1)	(3)	(4)	(4)-(3)
平均報酬率	0.0025	0.0062	0.0037 (0.0055)	0.0121	0.0119	-0.0002 (0.0096)
夏普比率	0.9170	1.1330	0.3960 (0.4276)	0.7665	0.8841	0.1176 (0.1174)
平均變異數	0.0000368	0.0000486	0.0000101 (0.0000101)	0.0000719	0.0000647	-4.45e-06 (0.0000142)

註: 1.括號內數字為估計標準誤; 2.***代表 p-value < 1%、**代表 p-value < 5%、*代表 p-value < 10%

5.1.3 模型資產配置 lead-lag 檢定

為進一步探討深度強化學習在納入總經資訊下的資產配置變化，本文分別檢視不同模型間的對應關係，以及同一模型於有無總經變數下的配置差異，結果整理如表 22 至表 24 所示。表 22 顯示，雖 DDQN 與 PPO 皆納入總經指標作為輔助資訊，但在資產配置上仍顯現差異。以 SPY 資產為例，DDQN 與 PPO（滯後 3 期）呈顯著負相關（r = -0.3321）；SHY 資產則在 PPO 領先 1 期與 4 期時，與 DDQN 分別呈顯著正相關（r = 0.4263 與 0.3998），顯示即便觀察空間與目標函數相同，兩模型決策邏輯仍具差異。表 23 進一步比較 DDQN 於有無總經指標下的表現，SHY 資產在有總經資訊下，相較無總經資訊，在領先 1 期與 4 期皆呈顯著正相關（r = 0.4265 與 0.3889），反映總經變數確實影響 DDQN 策略。表 24 則顯示，PPO 於有無總經條件下，對 DBC 資產在落後 5 期與領先 3 期皆呈顯著負相關（r = -0.3736 與 -0.3499），同樣顯示總經資訊對 PPO 決策有實質影響。

綜上而言，表 22 至 24 顯示，即使深度強化學習在觀察空間與訓練目標設定相似的情況下，DDQN 與 PPO 兩種策略於資產配置上仍展現出顯著異質性，特別在 SPY 與 SHY 資產中，呈現不同期數下的顯著正負相關。此外，同一策略在有無總經指標輔助下的表現亦具差異，特別是在 SHY (DDQN) 與 DBC (PPO) 資產的領先或滯後期中，呈現顯著不同的決策方向。因此，「是否納入總經指標」對於同一深度強化學習策略確實會造成配置行為的改變，反映出模型在整合總體資訊後對市場節奏的反應能力。然而，就深度強化學習與馬可維茲模型的策略差異而言，二者本質上即為不同決策邏輯與學習機制，總經變數難以作為解釋深度學習策略是否在投資配置上領先或落後於馬可維茲模型的依據。此結果進一步說明，總經指標的納入雖具潛在助益，但其效果高度依賴模型設計與策略架構。

表 22：PPO 與 DDQN 各期相關性（僅列顯著者）

資產	組合	相關係數
SPY	DDQN (當期) vs PPO (滯後 3 期)	-0.3221* (0.0826)
SHY	DDQN (當期) vs PPO (領先 1 期)	0.4263** (0.0150)
	DDQN (當期) vs PPO (領先 4 期)	0.3998** (0.0316)

註:1.括號內數字為 p 值; 2.***代表 p-value < 1%、**代表 p-value < 5%、*代表 p-value < 10%

表 23：DDQN 有無總經指標之各期相關性（僅列顯著者）

資產	組合	相關係數
SHY	有總經 (當期) vs 無總經 (領先 1 期)	0.4263** (0.015)
	有總經 (當期) vs 無總經 (領先 4 期)	0.3998** (0.0316)

註:1.括號內數字為 p 值; 2.***代表 p-value < 1%、**代表 p-value < 5%、*代表 p-value < 10%

表 24：PPO 有無總經指標之各期相關性（僅列顯著者）

資產	組合	相關係數
DBC	有總經 (當期) vs 無總經 (領先 5 期)	-0.3736* (0.0502)
DBC	有總經 (當期) vs 無總經 (落後 3 期)	-0.3499* (0.0580)

註:1.括號內數字為 p 值; 2.***代表 p-value < 1%、**代表 p-value < 5%、*代表 p-value < 10%

5.1.4 多變量迴歸分析結果

本研究進一步透過多變量迴歸分析，比較馬可維茲、 PPO 與 DDQN 模型於不同資產配置下對總體經濟變數的敏感程度（詳見表 25 與表 26）。整體結果顯示，馬可維茲在所有資產與總經之標內迴歸係數皆不顯著（詳見附錄 3），符合預期。本研究迴歸分析中，更著重 PPO 與 DDQN 模型之結果。而結果顯示 PPO 策略在納入總經變數後，僅對 SHY 顯示出與失業率之顯著正相關，展現對經濟衰退風險的初步辨識能力，然而在無總經變數的情境下，其資產配置行為與各項總經指標皆無顯著關聯，突顯總經資訊對於 PPO 學習過程與資產配置決策的重要性與必要性。值得注意的是，DDQN 展現出顯著且具經濟意涵的總經反應性，其中 IPI、ΔCPI 與 UMP 為關鍵驅動因子。

具體而言，在 DDQN 模型中 SPY 資產對失業率之迴歸係數為 -50，UMP 亦達顯著水準（係數約為 +58），RMTS 迴歸係數為 -1.8，Initial 雖趨近於 0，但仍具統計顯著性；SHY 資產中，IPI 的迴歸係數為 -15，失業率為 +45，皆達顯著；DBC 資產方面，ΔCPI 迴歸係數為 -6.4、失業率為 -11，CLI 與失業率亦顯著但迴歸係數趨近於 0。結果顯示，即便在未納入總經變數的設定下，DDQN 仍能透過市場報酬率與報價結構隱含學習總經節奏，顯示其 off-policy 架構與經驗回放記憶機制有助於捕捉景氣循環特徵。尤其在納入總經資訊情境下，DDQN 策略展現出更符合經濟理論與市場邏輯的資產配置行為，例如 SPY 對 IPI 顯著正向、對 UMP 顯著負向反應，SHY 對 UMP 顯著正向反應，顯示總經變數能有效引導模型做出方向正確且具一致性的決策。整體而言，SPY 與 SHY 回歸模型的 R^2 值分別為 0.49 與 0.45，進一步證實其資產配置具備對總經變數的高度敏感性與解釋力。

綜合而言，DDQN 優於 PPO 能有效整合總經指標進行策略調整，展現出高度的實務應用潛力與理論貢獻；而馬可維茲模型則僅透過歷史數據間接反映總經變動，調整能力有限。研究結果支持本文假設，深度強化學習模型（特別是 off-policy 架構）具備更佳的總經整合能力與策略穩健性，適合作為中長期資產配置的基礎工具。

表 25：PPO 與 DDQN 之權重多變數迴歸結果 (n=33)

	PPO			DDQN		
總經指標	SPY	SHY	DBC	SPY	SHY	DBC
ΔRMTS	4.2944	-3.8025	-0.4919	0.2679	-0.5368	-0.2689
	(6.3086)	(3.5607)	(4.7385)	(5.0692)	(4.2795)	(2.7254)
ΔIPI	-3.4873	5.3616	-1.8743	20.181***	19.3559***	0.8258
	(8.3881)	(4.7344)	(6.3004)	(6.7402)	(5.6902)	(3.6236)
ΔCPI	-13.2819	-10.6421	23.9240	21.799	-30.2029	8.4039
	(26.9828)	(15.2298)	(20.2673)	(21.6819)	(18.3043)	(11.6563)
CLI	-16.7353	-0.1110	0.6187	0.1330	-0.0830	-0.0500
	(0.1366)	(0.0771)	(0.1026)	(0.1098)	(0.0927)	(0.0590)
UMP	-2.18e-06	34.2289*	-17.4926	-49.893*	45.5255*	4.3677*
	(34.3883)	(19.4096)	(25.8296)	(27.6325)	(23.3279)	(14.8554)
Initails	-3.3836	1.98e-06	2.00e-07	4.85e-07	-3.72e-06	3.23e-06
	(3.69e-06)	(2.08e-06)	(2.77e-06)	(2.97e-06)	(2.50e-06)	(1.59e-06)
df	26	26	26	26	26	26
R^2	0.0566	0.2067	0.0770	0.3501	0.4493	0.1905

..註:1.括號內數字為估計標準誤; 2.***代表 p-value < 1%、**代表 p-value < 5%、*代表 p-value < 10%

表 26：PPO 與 DDQN 無總經指標之權重多變數迴歸結果 (n=33)

	PPO			DDQN		
總經指標	SPY	SHY	DBC	SPY	SHY	DBC
ΔRMTS	0.3400	-1.4669	1.1269	-1.8190***	1.0975	0.7216
	(4.0671)	(4.3322)	(4.9383)	(4.0524)	(4.0026)	(0.5776)
ΔIPI	6.7041	-9.7975	3.0034	15.2294	-15.2746***	0.0447
	(5.4078)	(5.7603)	(6.5662)	(5.3883)	(5.3220)	(0.7679)
ΔCPI	1.4922	20.2403	-21.7353	20.5974	-14.1820	-6.4154**
	(17.3957)	(18.5297)	(21.1220)	(17.3331)	(17.1199)	(2.4702)
CLI	-0.0486	-0.0753	0.1240	-0.1350	0.1068	0.0282**
	(0.0881)	(0.0938)	(0.1070)	(0.0878)	(0.0866)	(0.0125)
UMP	4.1641	-8.3266	4.1624	52.8628**	-41.9202	-10.9426**
	(22.1700)	(23.6153)	(26.9190)	(22.0901)	(21.8185)	(3.1482)
Initials	4.4822	-1.83e-06	-5.82e-08	4.18e-06*	-4.07e-06	-1.08e-07
	(8.0918)	(2.54e-06)	(2.89e-06)	(2.37e-06)	(2.34e-06)	(3.38e-07)
df	26	26	26	26	26	26
R^2	0.1059	0.1626	0.2814	0.4898	0.4546	0.4728

註: 1.括號內數字為估計標準誤; 2.***代表 p-value < 1%、**代表 p-value < 5%、*代表 p-value < 10%

第二節 結論與建議

5.2.1 結論

本研究以馬可維茲均值-變異數模型作為比較基準，其以歷史資料估計資產的期望報酬率與共變異數矩陣，進而求取最適投資組合權重以最大化夏普比率。然而，該模型高度仰賴靜態歷史數據，難以即時反映市場變化與景氣循環，且在整合總體經濟資訊方面亦存侷限，限制其於動態資產配置上的應用彈性。

為克服此一限制，本文進一步引入深度強化學習演算法中的 PPO 與 DDQN，藉由引入資產報酬、總體經濟變數等狀態資訊，訓練動態學習與調整的資產配置策略。PPO 採用連續型動作空間與夏普比率為獎勵函數，能在穩定學習過程中兼顧風險與報酬，展現出對高波動市場的良好適應能力；DDQN 則透過離散型動作空間搭配類似的狀態設定與獎勵函數，具備穩健策略學習能力與延遲調整特性，即使在無總經資訊輔助下亦能做出適應性資產配置。另一方面，實證亦顯示本研究所選取之六項總體經濟指標具備明顯的景氣循環區辨能力，特別是 ΔCPI、CLI、$\Delta RMTS$ 和 ΔIPI，於不同景氣階段呈現顯著差異。而在 10% 顯著水準下亦具區分能力，佐證其納入模型可提升資產配置策略的宏觀適應性與決策依據。

整體研究流程透過月資料進行模型訓練與測試，並從權重分配、報酬率與風險表現進行檢定，再以交叉相關與多變量迴歸等方法分析強化學習模型是否能在資產配置上展現出節奏性優勢與對總經指標的敏感性。

綜合三項研究問題，本文結論如下：

1. 深度強化學習模型能提升投資組合績效表現，能提升投資組合績效表現
   本研究實證結果顯示，DDQN 與 PPO 能夠根據市場與總經環境動態調整資產配置。PPO 在納入總經變數後展現更高的調整靈敏度，適合應對快速變動市場；DDQN 則展現穩健且具延遲反應特性，適用於中長期配置情境。與馬可維茲模型之靜態配置相比，即便馬可維茲加入滾動視窗，深度強化學習策略仍展現更高的平均報酬率與夏普比率。
2. off-policy 方法較 on-policy 方法更適合進行資產配置
   從績效表現來看，DDQN 策略於樣本期間展現出最佳的平均報酬率與夏普比率，顯示其能有效平衡報酬與風險，在不同資產間分配均衡，不易產生風險集中現象。相較之下，PPO 雖在整體績效略遜，但具備較強的策略調整能力，對總經變數高度敏感，具捕捉市場節奏的潛力。因此，off-policy 架構更適合應用於強調穩定性與風險控制的中長期投資策略。

3. 深度強化學習方法和馬可維茲模型沒有領先-落後關係
   雖然深度強化學習模型納入總經資訊後，能改善策略反應與適應性，且也和馬可維茲沒有領先-落後關係。由此可知，馬可維茲模型無法因應總體經濟變化。此現象可能來自兩類模型於決策邏輯與學習機制上的根本差異，無法單以時間節奏作為優劣判斷依據。因此，「是否納入總經指標」確實會影響強化學習策略行為，但尚無充分證據證實其具備明確的時間領先優勢。此外，根據多變量迴歸的結果顯示，即便在未納入總經變數的設定下，DDQN 仍能透過市場報酬與報價結構隱含學習總經節奏，顯示其 off-policy 架構與經驗回放記憶機制有助於捕捉景氣循環特徵。SPY 與 SHY 迴歸模型的 R^2 值分別為 0.49 與 0.45，進一步證實其資產配置具備對總經變數的高度敏感性與解釋力。

綜上所述，深度強化學習策略在報酬績效、配置靈活性與總經整合能力上均優於傳統馬可維茲模型，特別在複雜市場環境中展現出高度應用潛力。整體而言，深度學習模型雖未明確「領先」傳統方法，但其靈活學習與非線性決策特性，已為動態資產配置提供可行且具競爭力之策略方向。

5.2.2 研究建議

本研究雖針對深度強化學習與傳統資產配置模型進行多面向分析，仍存在以下限制與待補充方向：

1. 樣本期間與市場條件有限
   本研究使用 2022–2025 年之市場資料進行訓練與檢驗，期間受限於單一景氣循環階段，難以完整涵蓋如 Brocato & Steed (1998) 所述的擴張與衰退交替情境。未來可延伸樣本期間至歷經完整景氣週期，以提升模型於不同經濟環境下之穩健性與績效結果。
2. 總經變數與建模技術仍可擴充其他指標與資訊
   雖已參考 Dzikevičius & Vetrov (2013) 與 Koenig & Emery (1994) 的研究納入 IPI、CPI、CLI 等指標，然總體變數仍相對有限，且 CLI 存在滯後性問題。建議可進一步引入利率結構、貨幣供給、製造業訂單等領先性因子，並結合 LSTM、Transformer 等先進時間序列建模技術，強化總經節奏辨識與策略反應能力。
3. 研究資產數量較少，限制策略受限
   本研究以 SPY、SHY 與 DBC 三類資產為例，雖具代表性，惟仍未能捕捉如 Jensen & Mercer (2003) 與 Lu & Su (2011) 所指出在不同資產類型間景氣敏感度差異的全貌。建議未來擴展至多資產類別與跨市場（如 REITs、新興市場股市等），以驗證模型於多維風險環境下之適應性與資產輪動能力。

4. 獎勵設計仍可優化以提升模型能力
   本研究發現 DDQN 對總經變數具反應力但仍偏穩定，顯示現行獎勵設計仍未充分發揮總經指標預警功能。未來可結合 Vliet and Blitz (2011) 所提「不同階段風險變異顯著」的觀點，設計以「景氣階段動態權重」、「多目標最適化」為基礎之獎勵函數，提升模型辨識與調整能力。
5. 演算法選擇仍具有優化與其他選擇空間
   如第二章所述，多篇研究已證實不同強化學習架構（如 A3C、PPO、DDQN、Dueling DDQN、TD3、SAC）在資產配置任務中展現出異質性的學習特性與策略表現（Kim et al. 2019; Baek, 2024; Bajpai, 2021; 黃牧天, 2021; Bhardwaj et al. 2024）。例如，off-policy 演算法如 Dueling DDQN 或 SAC 在穩定報酬與控制回撤方面表現優異，而 on-policy 方法如 PPO 則較適合應對市場波動，展現出更高的決策穩健性。

總結而言，未來研究若能融合景氣循環理論、多維總經指標與先進演算法架構，並橫跨更長期與更多樣化市場條件，將可進一步驗證深度強化學習於資產配置任務中的穩健性與經濟意涵，回應傳統模型於估計誤差與靜態參數上的限制，強化其作為動態投資決策工具之應用潛力。

參考文獻

柯元富（2022）。〈Double DQN 模型應用於自動股票交易系統〉。未出版碩士論文，國立臺灣科技大學。

張愷凌（2009）。〈景氣循環、總體經濟變數與台灣股價指數的關係性研究〉。未出版碩士論文，國立交通大學。

黃冠棋（2021）。〈應用深度雙 Q 網路於股票自動交易系統〉。未出版碩士論文，國立政治大學。

黃牧天（2021）。〈應用深度強化學習演算法於資產配置優化之比較〉。未出版碩士論文，國立政治大學。

楊晴穎（2021）。〈深度強化學習於投資組合管理交易策略〉。未出版碩士論文，國立臺北大學。

謝妲娟（2012）。〈單根檢定與結構性改變─調查與運用〉。碩士論文，南華大學。

Abhishek, B., Krishi, K., Meghana, M., Daaniyaal, M., & Anupama, H. S. (2020), " Hand gesture recognition using machine learning algorithms," Computer Science and Information Technologies, 1(3), 116–120.

Aithal, P. K., Geetha, M., Savitha, B., & Menon, P. (2023), "Real-time portfolio management system utilizing machine learning techniques," IEEE Access, 11, 32595–32608.

Aruoba, S. B., Diebold, F. X., Kose, M. A., & Terrones, M. E. (2011), "Globalization, the business cycle, and macroeconomic monitoring," In NBER International Seminar on Macroeconomics 2010 (pp. 245–286). National Bureau of Economic Research.

Baek, S. M., et al. (2024), "Prediction model of browning inhibitor concentration and its optimal composition," Journal of Food Science, 89(8), 4986–4996.

Bajpai, S. (2021,. "Application of deep reinforcement learning for Indian stock trading automation," arXiv preprint arXiv:2106.16088.

Bartram, S. M., Branke, J., De Rossi, G., & Motahari, M. (2021), "Machine learning for active portfolio management," Journal of Financial Data Science, 3(3), 9–30.

Benhamou, E. (2023), "Can deep reinforcement learning solve the portfolio allocation problem? (Doctoral dissertation)," Université Paris sciences et lettres.

Bhardwaj, G. S., Pratap, D., & Darapaneni, N. (2024), "Optimized automated stock trading using DQN and Double DQN," In 2024 International Conference on Intelligent Algorithms for Computational Intelligence Systems (IACIS), IEEE.

Brocato, J., & Steed, S. (1998), "Optimal asset allocation over the business cycle," Financial Review, 33(3), 129–148.

Brown, D. B., & Smith, J. E. (2011), “Dynamic portfolio optimization with transaction costs: Heuristics and dual bounds,” Management Science, 57(10), 1752–1770.

Buehler, H., et al. (2019), “Deep hedging: Hedging derivatives under generic market frictions using reinforcement learning,” Swiss Finance Institute Research Paper, (No. 19-80).

Burns, A. F., & Mitchell, W. A. (1946), “Measuring business cycles,” National Bureau of Economic Research.

Chakravorty, G., Awasthi, A., & Da Silva, B. (2018), “Deep learning for global tactical asset allocation,” SSRN Electronic Journal.

Chen, W., Zhang, H., Mehlawat, M. K., & Jia, L. (2021), “Mean–variance portfolio optimization using machine learning-based stock price prediction,” Applied Soft Computing, 100, 106943.

Clifton, J., & Laber, E. (2020), “Q-learning: Theory and applications,” Annual Review of Statistics and Its Application, 7(1), 279–301.

Cont, R. (2001), “Empirical properties of asset returns: Stylized facts and statistical issues,” Quantitative Finance, 1(2), 223–236.

Corazza, M., & Sangalli, A. (2015), “Q-Learning and SARSA: A comparison between two intelligent stochastic control approaches,” Ca' Foscari University Working Paper Series.

DeMiguel, V., Garlappi, L., & Uppal, R. (2009), “Optimal versus naive diversification: How inefficient is the 1/N portfolio strategy?,” The Review of Financial Studies, 22(5), 1915–1953.

Dickey, D. A., & Fuller, W. A. (1979), “Distribution of the estimators for autoregressive time series with a unit root,” Journal of the American Statistical Association, 74(366a), 427–431.

Dorokhova, M., et al. (2021), “Deep reinforcement learning control of EV charging with solar power, “Applied Energy, 301, 117504.

Duan, Y., et al. (2016), “Benchmarking deep RL for continuous control,” In ICML (pp. 1329–1338). PMLR.

Durall, R. (2022), “Asset allocation: From Markowitz to deep reinforcement learning,” SSRN Electronic Journal.

Dzikevičius, A., & Vetrov, J. (2012), “Stock market analysis through business cycle approach” Business: Theory and Practice, 13(1), 36–42.

Dzikevičius, A., & Vetrov, J. (2013), “Investment portfolio management using the business cycle approach,” Business: Theory and Practice, 14(1), 57–63.

Edmonds, M., Tenenbaum, J. B., & Griffiths, T. L. (2018), "Human causal transfer: Challenges for deep reinforcement learning," In Proceedings of the Annual Meeting of the Cognitive Science Society, 40.

Enders, W. (2014). Applied Econometric Time Series (4th ed.)," John Wiley.

Evan, F. K., & Emery, K. M. (1994), "Why the composite index of leading indicators does not lead," Contemporary Economic Policy, 12(1), 52–66.

Fehrle, D. (2020), "Asset allocation and the business cycle"

Fossati, S. (2016), "Dating US business cycles with macro factors," Studies in Nonlinear Dynamics & Econometrics, 20(5), 529–547.

Gennotte, G. (1986), "Optimal portfolio choice under incomplete information," The Journal of Finance, 41(3), 733–746.

Gogas, P., & Papadimitriou, T. (2021), "Machine learning in economics and finance," Computational Economics, 57, 1–4.

Greene, W. (2012), "Econometric Analysis (7th ed.)," Prentice Hall.

Guidolin, M., & Timmermann, A. (2007), "Asset allocation under multivariate regime switching," Journal of Economic Dynamics and Control, 31(11), 3503–3544.

Guo, Q. (2022), "Review of research on markowitz model in portfolios," In 2022 7th International Conference on Social Sciences and Economic Development (ICSSED 2022) (pp. 786–790). Atlantis Press.

Haarnoja, T., Zhou, A., Abbeel, P., & Levine, S. (2018), "Soft Actor-Critic: Off-policy maximum entropy deep reinforcement learning with a stochastic Actor," In ICML (pp. 1861–1870). PMLR.

Hasselt, H. V. (2010), "Double Q-learning," In Advances in Neural Information Processing Systems, 23.

Hasselt, H. V., Guez, A., & Silver, D. (2016), "Deep reinforcement learning with double Q-learning," In Proceedings of the AAAI Conference on Artificial Intelligence, 30(1).

Hieu, L. T. (2020), "Deep reinforcement learning for stock portfolio optimization," arXiv preprint arXiv:2012.06325.

Huang, C. Y. (2018), "Financial trading as a game: A deep reinforcement learning approach," arXiv preprint arXiv:1807.02787.

Jensen, G. R., & Mercer, J. M. (2003), "New evidence on optimal asset allocation," Financial Review, 38(3), 435–454.

Jiang, T., Gradus, J. L., & Rosellini, A. J. (2020), "Supervised machine learning: a brief primer," Behavior Therapy, 51(5), 675–687.

Jin, Y., Qu, R., & Atkin, J. (2016), "Constrained portfolio optimisation: The state-of-the-art Markowitz models," In International Conference on Operations Research and Enterprise Systems, 2, 388–390.

Khare, I. S., Martheswaran, T. K., & Dassanaike-Perera, A. (2023), “Evaluation of reinforcement learning techniques for trading on a diverse portfolio,” arXiv preprint arXiv:2309.03202.

Khemlichi, F., Chougrad, H., Ali, S. E. B., & Khamlichi, Y. I. (2023), “Multi-Agent Proximal Policy Optimization for Portfolio Optimization,” Journal of Theoretical and Applied Information Technology, 101(20).

Kim, J. B., Heo, J. S., Lim, H. K., Kwon, D. H., & Han, Y. H. (2019), “Blockchain Based Financial Portfolio Management Using A3C,” KIPS Transactions on Computer and Communication Systems, 8(1), 17–28.

Kinlaw, W., Kritzman, M., & Turkington, D. (2021), “A new index of the business cycle,” Journal of Investment Management, 19(3), 4–19.

Lei, X., et al. (2022), “Development of an intelligent information system for financial analysis depend on supervised machine learning algorithms,” Information Processing & Management, 59(5), 103036.

Liang, Z., et al. (2018), “Adversarial deep reinforcement learning in portfolio management,” arXiv preprint arXiv:1808.09940.

Lu, Y., & Su, M. (2011), “Asset allocation model across business cycle,” In 2011 International Conference on Business Management and Electronic Information, 2, 327–330. IEEE.

Markowitz, H. (1952), “Portfolio Selection,” The Journal of Finance, 7(1), 77–91.

McKenziea, A., & McDonnell, M. D. (2023), “Hyperparameter selection in reinforcement learning using the “design of experiments” method,” Procedia Computer Science, 222, 11–24.

Michaud, R. O. (1989), “The Markowitz optimization enigma: Is ‘optimized’ optimal?,” Financial Analysts Journal, 45(1), 31–42.

Mnih, V., et al. (2015), “Human-level control through deep reinforcement learning,” Nature, 518(7540), 529–533.

Mnih, V., et al. (2016), “Asynchronous methods for deep reinforcement learning,” In ICML (pp. 1928–1937). PMLR.

Obeidat, M. A., et al. (2021), “A deep review and analysis of artificial neural network use in power application,” In 2021 International Renewable Engineering Conference (IREC), 1–5. IEEE.

Pendharkar, P. C., & Cusatis, P. (2018), “Trading financial indices with reinforcement learning agents,” Expert Systems with Applications, 103, 1–13.

Pokou, F., et al. (2025), “Bridging Econometrics and AI: VaR Estimation via Reinforcement Learning and GARCH Models,” arXiv preprint arXiv:2504.16635.

Schulman, J., et al. (2017), “Proximal policy optimization algorithms,” arXiv preprint arXiv:1707.06347.

Sen, J., Mehtab, S., & Dutta, A. (2020), “Stock price prediction using machine learning and LSTM,” In Symposium on Machine Learning and Metaheuristics, 88–106.

Sengupta, J. K. (1989), “A dynamic view of the portfolio efficiency frontier,” Computers & Mathematics with Applications, 18(6–7), 565–580.

Sood, S., et al. (2023), “Deep reinforcement learning for optimal portfolio allocation: A comparative study,” FinPlan, 2023(2023), 21.

Sutton, R. S., & Barto, A. G. (1998), “The reinforcement learning problem,” Reinforcement Learning: An Introduction, 51–85.

Usman, A. (2023), “Algorithmic Alpha: A Comparative Analysis of Machine Learning and Classical Approaches for Stock Portfolio Optimisation,” SSRN 4994831.

Van Vliet, Pim and Blitz, David and van der Grient, Bart(2011), “Is the Relation between Volatility and Expected Stock Returns Positive, Flat or Negative?,” Available at SSRN: http://dx.doi.org/10.2139/ssrn.1881503

Wang, H. (2024), “Constrained Portfolio Optimization: Markowitz Model and Index Model,” In SHS Web of Conferences, 208, 04021. EDP Sciences.

Watkins, C. J., & Dayan, P. (1992), “Q-learning,” Machine Learning, 8, 279–292.

Wongsawatgul, B. (2019), “An Investigation of ETF Flows: Asset Allocation Perspectives.”

Yan, Z., et al. (2024), “Inter-layer feedback mechanism with reinforcement learning boosts the evolution of cooperation,” Chaos, Solitons & Fractals, 185, 115095.

Yang, H., Park, H., & Lee, K. (2022), “A Selective Portfolio Management Algorithm with Off-Policy Reinforcement Learning Using Dirichlet Distribution,” Axioms, 11(12), 664.

Yang, H., et al. (2020), “Deep reinforcement learning for automated stock trading: An ensemble strategy,” In ACM AI in Finance, 1–8.

Zarnowitz, V. (1992), “What is a business cycle? ,”In The Business Cycle: Theories and Evidence, 3–83.

Zhang, Y., & Ross, K. W. (2021), “On-policy deep reinforcement learning for the average-reward criterion,” In ICML, 12535–12545.

Zhang, Y., Zohren, S., & Roberts, S. (2020), “Deep reinforcement learning for trading,” arXiv preprint. https://arxiv.org/pdf/1911.10107

附錄一 中英名詞對照表

中文名詞	英文名詞	備註
A-		–
動作、行動	Action	-
演員–評論家	Actor-Critic	-
赤池資訊量準則	AIC	Akaike information criterion
優勢動作評論	A2C	Advantage Actor Critic
異步優勢演員–評論家	A3C	Asynchronous Advantage Actor-Critic
代理人	Agent	-
B-		-
批次大小	batch size	-
景氣循環	Business Cycles	-
C-		-
資本市場線	CML	Capital Market Line
綜合領先指標	CLI	Composite Leading Indicator
剪裁機制	Clipping	-
剪裁目標函數	clipped objective function	-
剪裁函數	clipping parameter	ϵ
卷積神經網路	CNN	Convolutional neural networks
限制最佳化	Constrained optimization	有拘束最佳化
連續型	Continuous	-
收縮期	Contraction	-
消費者物價指數	CPI	Consumer Price Index
數學程序求解器	CPLEX	-
價值函數	Critic	-
交叉相關函數	Cross-Correlation Function	-
D-	-	-
數據驅動	Data-driven	-
數據洩漏	Data-Leakage	-
深度確定性策略梯度	DDPG	Deep Deterministic Policy Gradient
雙重深度 Q 網路	DDQN	Double Deep Q Network

中文名詞	英文名詞	備註
		Double DQN
深度	Depth	-
折扣因子	discount factor	γ
離散型	discrete	-
競爭架構雙重深度 Q 網路	Dueling DDQN	Dueling Double Deep Q Network
深度強化學習技術	DRL	Deep Reinforcement Learning
夏普比率差值	Differential Sharpe Ratio	-
廣度	Diffusion	-
多角化	Diversification	-
持續時間	Duration	-
E-		-
效率前緣	Efficient Frontier	-
熵係數	Entropy Coefficient	-
環境	Environment	-
均分策略	Equal weight	-
期望效用理論	EUT	Expected Utility Theory
擴張期	Expansion	-
探索	Exploitation	-
利用	Exploration	-
梯度爆炸	Exploding gradient	-
F-		-
聯邦儲備經濟數據	FRED	Federal Reserve Economic Data
向前傳播	Forward Propagation	-
G-		-
國內生產毛額	GDP	Gross Domestic Product
全域最小風險投資組合	GMV	Global Minimum Variance Portfolio
網格搜索法	Grid search	-
I-		
指數模型	Index Model	-
初領失業救濟金人數	Initial Claims	-
工業生產指數	IPI	Industrial Production Index

中文名詞	英文名詞	備註
L-		
學習率	learning rate	-
水準定態	level stationary	-
長短期記憶模型	LSTM	Long Short-Term Memory
M-		
機器學習技術	ML	Machine Learning
總體經濟指標數據	Macro Variables	-
馬可維茲模型	Markowitz	Mean-Variance Portfolio
馬可維茲最適化	Markowitz Optimization	-
最大夏普比率投資組合	MSR	Maximum Sharpe Ratio Portfolio
最大回撤	MDD	Maximum Drawdown
拉格朗日乘數法	Method of Lagrange Multiplier	-
現代投資組合裡論	MPT	Modern Portfolio Theory
多輸入方法	Multi-input Method	-
N-		
美國國家經濟研究局	NBER	National Bureau of Economic Research
O-	-	-
觀測	Observation	-
經濟合作暨發展組織	OECD	Organization for Economic Cooperation and Development
異策略演算法	Off-policy	-
主網路	Online network	-
同策略演算法	On-policy	-
高估偏誤	Overestimation bias	-
P-		
皮爾森相關係數	Pearson Correlation Coefficient	-
投資組合權重	Portfolio Weights	-
策略梯度	Policy Gradient	-

中文名詞	英文名詞	備註
策略學習	Policy Learning	-
近端策略優化	PPO	Proximal Policy Optimization
Q-		
Q 學習	Q-learning	-
R-		
風險趨避行為者	Rational Risk Averter	-
實質製造業與貿易銷售額	Real Manufacturing and Trade Industries Sales	-
衰退期	Recession	-
經驗回放	replay memory	-
經驗回放記憶體	Replay memory size	-
獎勵函數	Reward Function	-
復甦期	Recovery	-
滾動視窗	Rolling Window	-
S-		
柔性動作-評論家	SAC	Soft Actor-Critic
狀態-動作-獎勵-動作算法	SARSA	State-Action-Reward-State-Action
夏普比率	Sharpe ratio	-
步長	step size	-
隨機梯度下降	Stochastic gradient descent	SGD
T-		
目標網路	Target network	-
雙延遲深度確定性策略梯度	TD3	Twin Delayed Deep Deterministic policy gradient
時間差分控制	TD Control	-
變換器	Transformer	-
信賴區間策略最佳化演算法	TRPO	Trust Region Policy Optimization
U-		
失業率	Unemployment rate	-
V-		
梯度消失	vanishing gradient	-
W-		

中文名詞	英文名詞	備註
觀察視窗	window size	-
X-		
極限梯度提升	XGBoost	Extreme gradient boosting

附錄二 總經指標雪費事後檢定

總經指標	景氣階段 (i)	景氣階段 (j)	平均差異 (i-j) (p 值)
ΔCPI	擴張期	衰退期	0.0016** (0.0418)
	回落期	復甦期	0.0016** (0.0215)
	回落期	衰退期	0.0029*** (<0.0001)
CLI	擴張期	復甦期	1.8539*** (<0.0001)
	擴張期	衰退期	2.1865*** (<0.0001)
	回落期	復甦期	1.8784*** (<0.0001)
	回落期	衰退期	2.2110*** (<0.0001)
$\Delta RMTS.$	擴張期	衰退期	0.0070** (0.0332)
	回落期	復甦期	-0.0057* (0.0983)
	復甦期	衰退期	0.0096*** (0.0005)
ΔIPI	擴張期	衰退期	0.0088*** (0.0016)
	回落期	衰退期	0.0065* (0.0511)
	復甦期	衰退期	0.0097*** (0.0002)

總經指標	景氣階段 (i)	景氣階段 (j)	平均差異 (i-j) (p 值)
Unemp.	擴張期	復甦期	-0.0093* (0.0624)
	回落期	復甦期	-0.0163*** (0.0003)
Initials	回落期	復甦期	-172564.00* (0.0662)

註:* 表示顯著水準 $p < 0.1$；
** 表示 $p < 0.05$；
*** 表示 $p < 0.001$

附錄三 馬可維茲多變量回歸分析結果

總經指標	馬可維茲		
	SPY	SHY	DBC
ΔRMTS	-3.6652	1.3446	2.2306
	(6.6008)	(9.5176)	(9.3694)
ΔIPI	2.0500	0.7689	-2.8190
	(8.7767)	(12.6549)	(12.4580)
ΔCPI	40.5741	-31.749	-8.8092
	(28.2329)	(40.7085)	(40.0749)
CLI	-0.0771	0.1658	-0.0886
	(0.1430)	(0.2061)	(0.2029)
UMP	-19.6367	-108.2098	127.8466
	(35.9815)	(51.8810)	(51.0735)
Initials	1.69e-06	-1.29e-05	-3.98e-06
	(3.86e-06)	(-5.57e-06)	(5.48e-06)
df	26	26	26
R2	0.4667	0.4403	0.4293

註: 1.括號內數字為估計標準誤; 2.***代表 p-value < 1%、**代表 p-value < 5%、*代表 p-value < 10%