Final Exam

學號： M11218014
姓名： 王士誠

---

1. 題目一：

When using time series methods to build a stock price prediction model, researchers can choose to directly use the stock price series as input or convert the price series into a daily return series before modeling. Please compare the theoretical basis, data characteristics, and model performance of these two methods in detail.

1.1 理論基礎比較

1.1.1 股價序列 (Price Series)

- **理論基礎**：
  - 基於隨機漫步理論 (Random Walk Theory)
  - 假設股價遵循幾何布朗運動 (Geometric Brownian Motion)
  - 股價具有趨勢性和持續性 (Trend and Momentum)
  - 符合有效市場假說中的弱式效率
- **數學表達**：

  $P_t = P_{t-1} + \varepsilon_t$

  或

  $P_t = \mu + \phi P_{t-1} + \varepsilon_t$

1.1.2 日報酬率序列 (Daily Return Series)

- **理論基礎**：
  - 基於對數正態分布假設
  - 符合現代投資組合理論 (Modern Portfolio Theory)
  - 報酬率具有較佳的統計特性（平穩性、常態性）
  - 更貼近實際投資決策思維
- **數學表達**：

  $R_t = \ln\left(\frac{P_t}{P_{t-1}}\right) = \ln(P_t) - \ln(P_{t-1})$

  或

  $R_t = \frac{P_t - P_{t-1}}{P_{t-1}}$

1.2 資料特性比較

1.2.1 股價序列特性

1. **非平穩性 (Non-stationarity)**：
   - 具有單根 (Unit Root)
   - 變異數隨時間增加
   - 均值可能存在結構性變化

2. **自相關性**：
    - 高度自相關
    - 長期記憶效應
    - 趨勢成分明顯
3. **分布特性**：
    - 通常不符合常態分布
    - 具有厚尾特徵 (Fat Tails)
    - 可能存在偏態 (Skewness)

1.2.2 日報酬率序列特性

1. **平穩性 (Stationarity)**：
    - 較接近弱平穩過程
    - 均值相對穩定
    - 變異數相對恆定
2. **自相關性**：
    - 低度自相關或無自相關
    - 符合白噪音假設
    - 更適合統計建模
3. **分布特性**：
    - 較接近常態分布
    - 仍具有波動度聚集現象 (Volatility Clustering)
    - 厚尾特徵依然存在但較股價序列溫和

1.3 模型表現比較

1.3.1 股價序列建模

優點：

- 直接預測目標價格，解釋性強
- 保留了價格水準資訊
- 適合長期趨勢預測

缺點：

- 非平穩性導致虛假迴歸 (Spurious Regression)
- 需要進行單根檢定和共整合分析
- 模型參數可能不穩定
- 預測誤差可能隨時間放大

適用模型：

- ARIMA 模型（需先進行差分）
- VAR 模型（向量自迴歸）
- 向量誤差修正模型（Vector Error Correction Model, VECM）

1.3.2 日報酬率序列建模

優點：

- 資料具有較佳的統計特性
- 避免虛假迴歸問題
- 模型參數較穩定
- 風險控制更精確

缺點：

- 無法直接預測價格水準
- 需要透過累積報酬計算價格
- 預測誤差會累積傳遞
- 可能忽略長期均值迴歸現象

適用模型：

- AR/MA 模型
- GARCH 族模型（處理波動度聚集）
- 機器學習模型（SVM, Random Forest, Neural Networks）

1.4 實證建議

根據文獻和實務經驗：

1. **短期預測 (1-5天)：** 建議使用日報酬率序列
    - 資料平穩性佳
    - 統計檢驗力強
    - 適合高頻交易策略
2. **中長期預測 (1週-1月)：** 可考慮股價序列
    - 趨勢資訊更完整
    - 但需注意共整合關係
    - 可結合基本面分析
3. **風險管理：** 優先使用報酬率序列
    - VaR 計算更準確
    - 波動度模型更適用
    - 符合現代風險管理框架

---

2. 題目二：

Is the relationship between option prices and their underlying asset prices linear? Based on theoretical foundations, explain the characteristics of the relationship between option prices and underlying asset prices. What statistical methods or empirical analysis strategies can verify this nonlinearity? Considering that the underlying asset price is one of the essential inputs, what algorithm would you recommend to model option prices? Explain your reasons for recommending the algorithm.

2.1 理論基礎：選擇權價格與標的資產價格關係

2.1.1 非線性關係的理論依據

選擇權價格與標的資產價格之間的關係**並非線性**，這可以從 Black-Scholes 模型清楚看出：

歐式買權的 Black-Scholes 公式：

$$C = S_0 e^{-q\tau}\Phi(d_1) - Ke^{-r\tau}\Phi(d_2)$$

其中：

$$d_1 = \frac{\ln(S_0/K) + (r - q + \sigma^2/2)\tau}{\sigma\sqrt{\tau}}$$

$$d_2 = d_1 - \sigma\sqrt{\tau}$$

2.1.2 關係特徵分析

1. **凸性 (Convexity)**：
    - 選擇權價格對標的價格的二階導數 (Gamma) 恆為正
    - 呈現凸函式特性
    - 標的價格變化對選擇權價格的影響非線性遞增
2. **Delta 特性**：
    - $\Delta = \frac{\partial C}{\partial S}$ 代表價格敏感度
    - Delta 隨標的價格變化而變化 $(0 < \Delta < 1 \text{ for calls})$
    - 價內程度越高，Delta 越接近 1
    - 價外程度越高，Delta 越接近 0
3. **Moneyness 效應**：
    - 價內 (In-the-Money)：選擇權價值主要由內在價值決定
    - 價外 (Out-of-the-Money)：選擇權價值主要由時間價值決定
    - 價平 (At-the-Money)：最大的 Gamma 值，對標的價格最敏感
4. **時間衰減 (Theta)**：
    - 隨著到期日接近，時間價值遞減
    - 價平選擇權的時間衰減最快
    - 深度價內或價外選擇權時間衰減較慢

2.2 非線性驗證的統計方法

2.2.1 圖形分析法

透過散布圖分析選擇權價格與標的資產價格的關係，並進行殘差分析檢查線性模型的適用性。若存在明顯的曲線型態或殘差呈現系統性偏差，則表示非線性關係。

2.2.2 統計檢驗法

- **Rainbow Test**： 檢驗線性 vs 非線性關係
- **RESET Test**： Ramsey 迴歸模型檢定
- **White Test**： 異質變異數檢定
- **BDS Test**： 檢驗時間序列的非線性相關性

2.2.3 距離相關係數 (Distance Correlation)

距離相關係數能夠檢測非線性相關性，值介於 0 和 1 之間，0 表示統計獨立，1 表示完全相關。相較於傳統 Pearson 相關係數僅能捕捉線性關係，距離相關係數可以識別任何型態的依賴關係。

2.3.4 互資訊分析 (Mutual Information)

基於資訊理論的互資訊可以量化兩個變數之間的相互依賴程度，特別適合檢測非線性關係。可透過 KNN 估計法計算連續變數的互資訊值。

2.3.5 模型比較法

比較線性模型與多項式模型的解釋力（R^2 值），若多項式模型顯著優於線性模型，則證實非線性關係的存在。

2.3.6 非參數檢驗

- **Spearman 相關係數**： 捕捉單調非線性關係
- **Kendall's tau**： 基於等級的相關性測度

2.3 推薦演算法：神經網路 (Neural Networks)

2.3.1 推薦理由

1. **非線性建模能力**：
    - 多層感知機可以逼近任意連續函式 (Universal Approximation Theorem)
    - 自動學習複雜的非線性關係
    - 無需事先假設函式形式
2. **多維輸入處理**：
    - 能夠同時處理標的價格、履約價格、到期時間、無風險利率、波動率、股利殖利率等多個輸入變數
    - 自動學習變數間的交互作用
3. **自動特徵工程**：
    - 隱藏層自動組合特徵
    - 學習複雜的交互作用
    - 捕捉高階項效應

2.3.2 架構設計建議

建議採用多層前饋神經網路，包含批量標準化 (BatchNormalization) 和 Dropout 層以防止過度配適。輸出層使用線性激勵函數進行選擇權價格預測。優化器推薦使用 Adam，損失函數採用均方誤差 (MSE)。

2.3.3 替代演算法考慮

1. **隨機森林 (Random Forest)**：
    - 處理非線性關係能力強
    - 對異常值較不敏感
    - 提供特徵重要性分析
    - 但可能在價格邊界處表現較差
2. **支援向量機 (SVM) with RBF kernel**：
    - 非線性映射能力強
    - 理論基礎扎實
    - 但計算複雜度較高
    - 超參數調整較困難
3. **梯度提升樹 (Gradient Boosting)**：
    - XGBoost/LightGBM/CatBoost 系列
    - 序列學習，逐步修正誤差
    - 實務表現優異
    - 但易過度配適

2.3.4 模型驗證策略

建議採用時間序列分割驗證 (TimeSeriesSplit) 以避免前瞻偏誤，並使用滾動窗口方法評估模型在不同市場環境下的穩定性。

---

3. 題目三：

In stock price or return prediction, models often utilize multi-dimensional features (such as technical indicators, fundamental data, and macroeconomic variables), which may be highly correlated (resulting in multicollinearity) and impact model performance. Please explain the impact of high correlation among multi-dimensional features and propose methods for utilizing machine learning algorithms to address this issue.

3.1 多元共線性的影響

3.1.1 統計推論問題

- **係數不穩定：** 微小的資料變化導致參數估計大幅波動
- **標準誤膨脹：** 迴歸係數的標準誤增大，t 統計量下降
- **統計顯著性下降：** 個別變數的顯著性檢驗失效
- **信賴區間擴大：** 參數估計的不確定性增加

3.1.2 模型解釋性問題

- **參數解釋困難：** 難以判斷個別特徵的真實影響
- **符號反轉：** 迴歸係數可能出現不合理的符號
- **經濟意義喪失：** 模型失去經濟解釋能力

3.1.3 預測性能問題

- **過度配適：** 模型在訓練集表現好，測試集表現差
- **泛化能力差：** 對新資料的預測能力下降
- **數值不穩定：** 矩陣接近奇異，計算誤差放大

3.2 多元共線性檢測方法

3.2.1 相關係數矩陣

計算特徵間的 Pearson 相關係數，並透過熱力圖視覺化。一般認為絕對值大於 0.7 的相關係數表示高度相關。

3.2.2 變異數膨脹因子 (VIF)

VIF 衡量某個特徵與其他特徵的線性關係程度：

- $VIF < 5$：輕微共線性
- $5 \leq VIF < 10$：中度共線性
- $VIF \geq 10$：嚴重共線性

3.2.3 條件指數 (Condition Index)

透過奇異值分解計算設計矩陣的條件指數，大於 30 表示嚴重共線性問題。

3.3 機器學習解決方案

3.3.1 降維技術

- **主成分分析 (PCA)**

  將原始特徵投影到低維空間，保留主要變異數來源。建議保留 95% 的累積變異數，能有效消除共線性同時保持資訊完整性。

- **獨立成分分析 (ICA)**

  適用於非高斯分布的訊號分離，能夠識別獨立的潛在因子。

- **t-SNE**

  非線性降維技術，適合資料視覺化和探索性分析。

3.3.2 特徵選擇方法

- **過濾式方法 (Filter Methods)**
  - **單變量選擇：** 基於 F 統計量或互資訊選擇重要特徵
  - **相關性過濾：** 移除高度相關的冗餘特徵
- **包裝式方法 (Wrapper Methods)**
  - **遞迴特徵消除 (RFE)：** 遞迴移除不重要特徵
  - **前向/後向選擇：** 逐步增加或移除特徵
- **嵌入式方法 (Embedded Methods)**
  - **LASSO 正則化：** L1 懲罰項自動進行特徵選擇
  - **彈性網路：** 結合 L1 和 L2 正則化的優點

3.3.3 正則化方法

- **Ridge 迴歸**

  透過 L2 正則化項控制參數大小，減緩共線性影響，但不會完全移除特徵。

- **LASSO 迴歸**

  L1 正則化能將不重要特徵的係數壓縮至零，自動進行特徵選擇。

- **Elastic Net**

  結合 Ridge 和 LASSO 的優勢，在群組相關特徵中選擇代表性特徵。

3.3.4 樹基模型方法

- **隨機森林**

  透過特徵重要性排名識別關鍵變數，對共線性相對不敏感，但仍建議預處理以提升解釋性。

- **梯度提升樹**

  XGBoost、LightGBM 等演算法內建正則化機制，能自動處理特徵間的相關性。

3.4 實務建議

3.4.1 檢測階段

- **探索性資料分析：** 先瞭解資料分布和相關性結構
- **多重檢測：** 同時使用相關係數、VIF、條件指數等方法
- **分群分析：** 識別高度相關的特徵群組

3.4.2 處理策略選擇

- **輕微共線性** $(VIF < 5)$**：** 使用 Ridge 或 Elastic Net 正則化
- **中度共線性** $(5 \leq VIF < 10)$**：** 結合特徵選擇和正則化
- **嚴重共線性** $(VIF \geq 10)$**：** 優先考慮降維技術或 VIF 剔除

3.4.3 模型選擇考量

- **線性模型：** 對共線性敏感，需要前處理
- **樹基模型：** 對共線性較不敏感，但仍建議處理
- **神經網路：** 可透過 Dropout 和正則化緩解問題

3.4.4 驗證評估策略

採用交叉驗證比較不同處理方法的效果，評估指標包括預測準確度、模型穩定性和解釋性。建議建立整合式的共線性處理管道，能根據資料特性自動選擇最適方法。

---

4. 結論

本次期末考試深入探討了機器學習在金融應用中的三個關鍵議題：

4.1 主要貢獻與發現

1. **時間序列建模方法論**
    - 股價序列與報酬率序列各有優劣，需根據預測目標和時間範圍選擇
    - 短期預測偏好報酬率序列，長期預測可考慮股價序列
    - 風險管理應用以報酬率序列為主
2. **非線性關係檢測與建模**
    - 選擇權定價展現明顯的非線性特徵，傳統線性方法不適用
    - 距離相關係數和互資訊是檢測非線性關係的有效工具
    - 神經網路在捕捉複雜非線性關係方面具有明顯優勢
3. **多元共線性處理策略**
    - 多元共線性是金融建模中的常見問題，需要系統性解決方案
    - 不同程度的共線性需要不同的處理策略
    - 整合多種方法可以達到最佳效果

4.2 實務應用價值

這些理論和方法在實際金融科技應用中具有重要價值：

- **量化交易系統：** 適當的時間序列處理和特徵選擇可提升策略績效
- **風險管理系統：** 非線性建模技術可更準確地評估衍生性商品風險
- **投資組合管理：** 多元共線性處理有助於構建更穩健的投資模型

4.3 未來研究方向

1. **深度學習應用：** 探索 Transformer、LSTM 等深度學習架構在金融預測中的應用
2. **多模態融合：** 結合文字、圖像、時間序列等多種資料型態
3. **可解釋性研究：** 開發更好的模型解釋技術，滿足金融監管需求