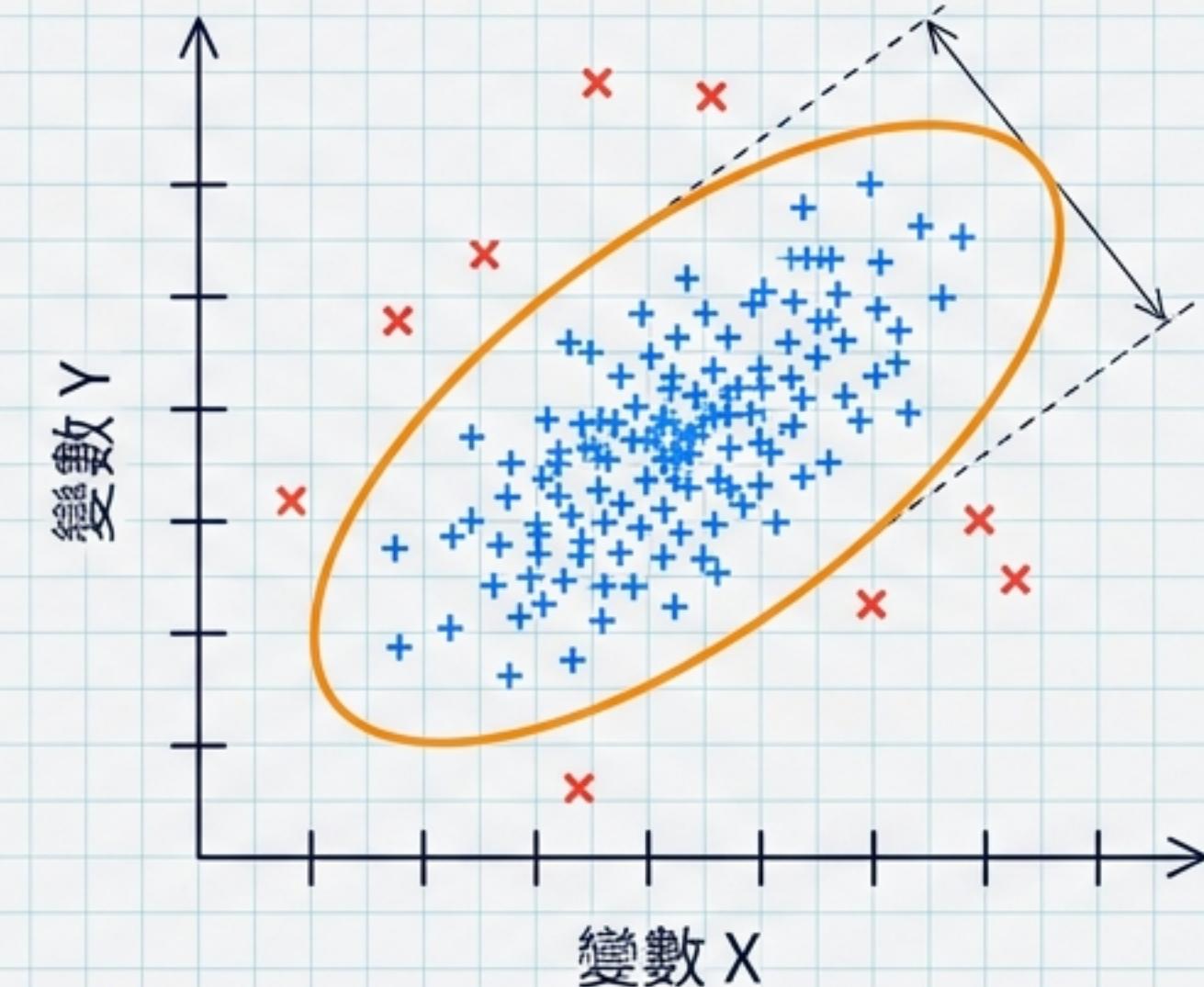


Unit 07: 楕圓包絡 (Elliptic Envelope)

基於高斯分布的異常檢測 (Gaussian-Based Anomaly Detection)



授課教師：莊曜禎 助理教授

課程目標：建立化工製程的「正常操作視窗」

定位：非監督式學習 (Positioning)

核心任務 (Core Task) :

在沒有標籤的情況下發現異常 (Discovering anomalies without labels)。

概念隱喻 (Metaphor) :

「大海撈針之前，先定義大海的形狀」

輸入 (Input) :
只有特徵數據 X

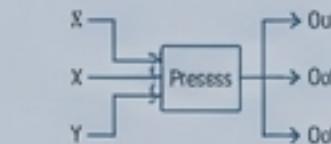
目標 (Goal) :
識別不屬於正常分布的孤立點

機器學習地圖：如何選擇正確的工具？

監督式學習
(Supervised Learning)



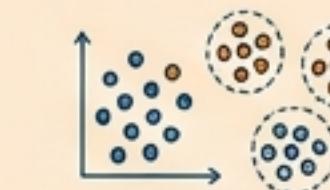
數據 (Data)：
有標籤數據 (Input X + Output y)
目標 (Goal)：
預測 (Prediction)
化工案例 (ChemE Examples)：
品質預測 (Quality Prediction)、
故障分類 (Fault Classification)



非監督式學習
(Unsupervised Learning)



數據 (Data)：
無標籤數據 (Input X only)
目標 (Goal)：
發現模式 (Discovery)
化工案例 (ChemE Examples)：
異常檢測 (Anomaly Detection)、
操作模式識別



強化學習
(Reinforcement Learning)



數據 (Data)：
環境互動 (Action/Reward)
目標 (Goal)：
決策優化 (Optimization)
化工案例 (ChemE Examples)：
製程控制策略 (Process Control)



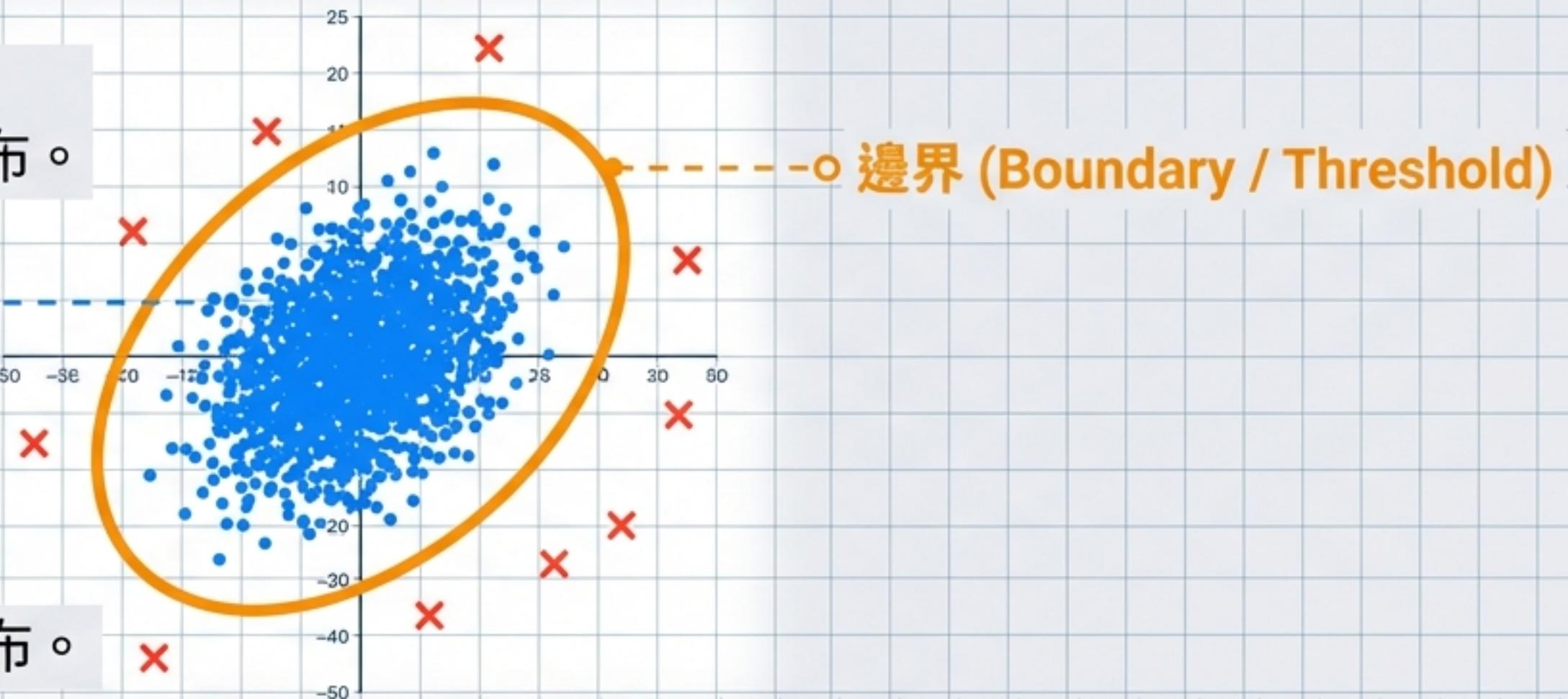
AI

什麼是橢圓包絡？(What is Elliptic Envelope?)

定義 (Definition) :

假設正常數據服從多變量高斯分布。

正常 (Normal / Center) ○ -----



定義 (Definition) :

假設正常數據服從多變量高斯分布。

核心假設 (Core Assumption) :

1. 正常數據聚集在分布中心 (Center)
2. 異常點是遠離中心的離群值 (Outliers)

為什麼假設高斯分布？(Why Gaussian?)

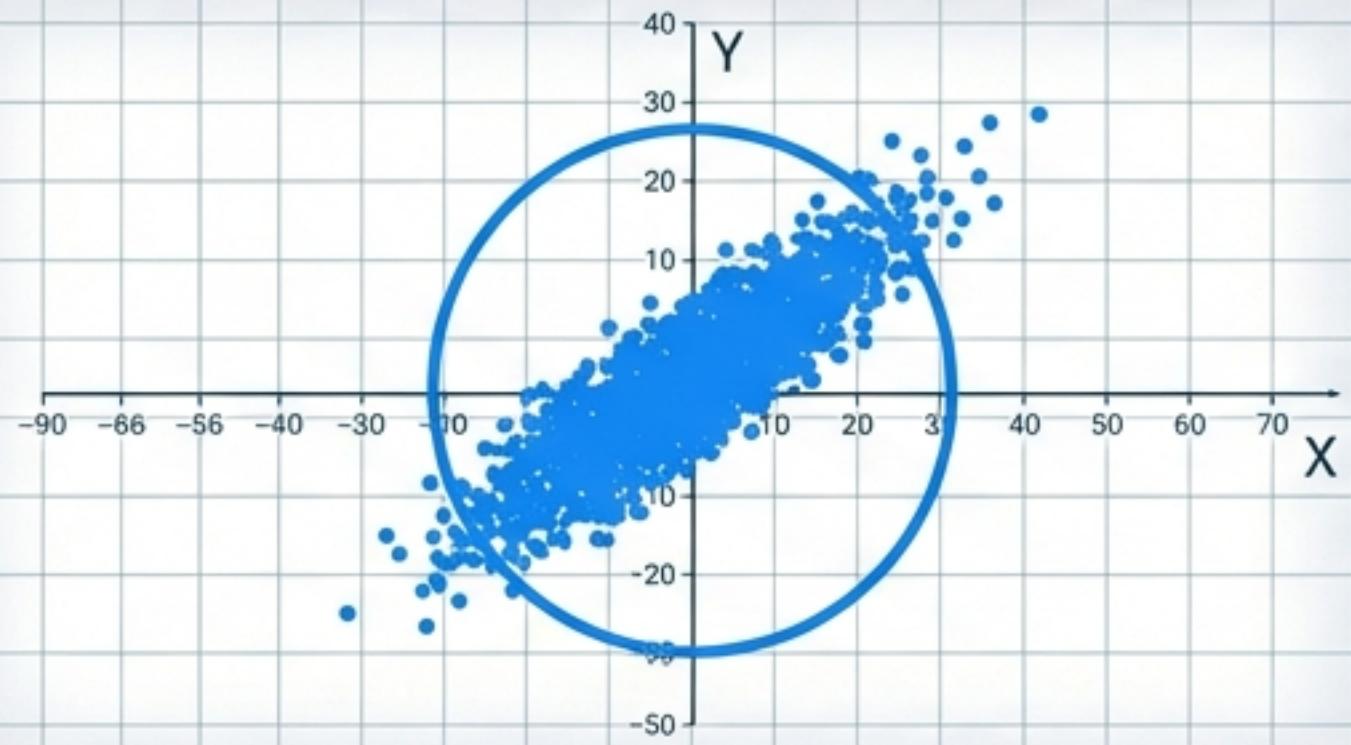


- 製程穩定性 (Process Stability)：良好控制的製程變數通常圍繞設定點波動。
- 感測器誤差 (Sensor Noise)：測量誤差通常服從常態分布。
- 中央極限定理 (Central Limit Theorem)：多個獨立隨機因素疊加的結果趨向常態分布。

結論：如果製程是穩定的，數據通常是高斯的。這使得橢圓包絡成為物理上正確的選擇。

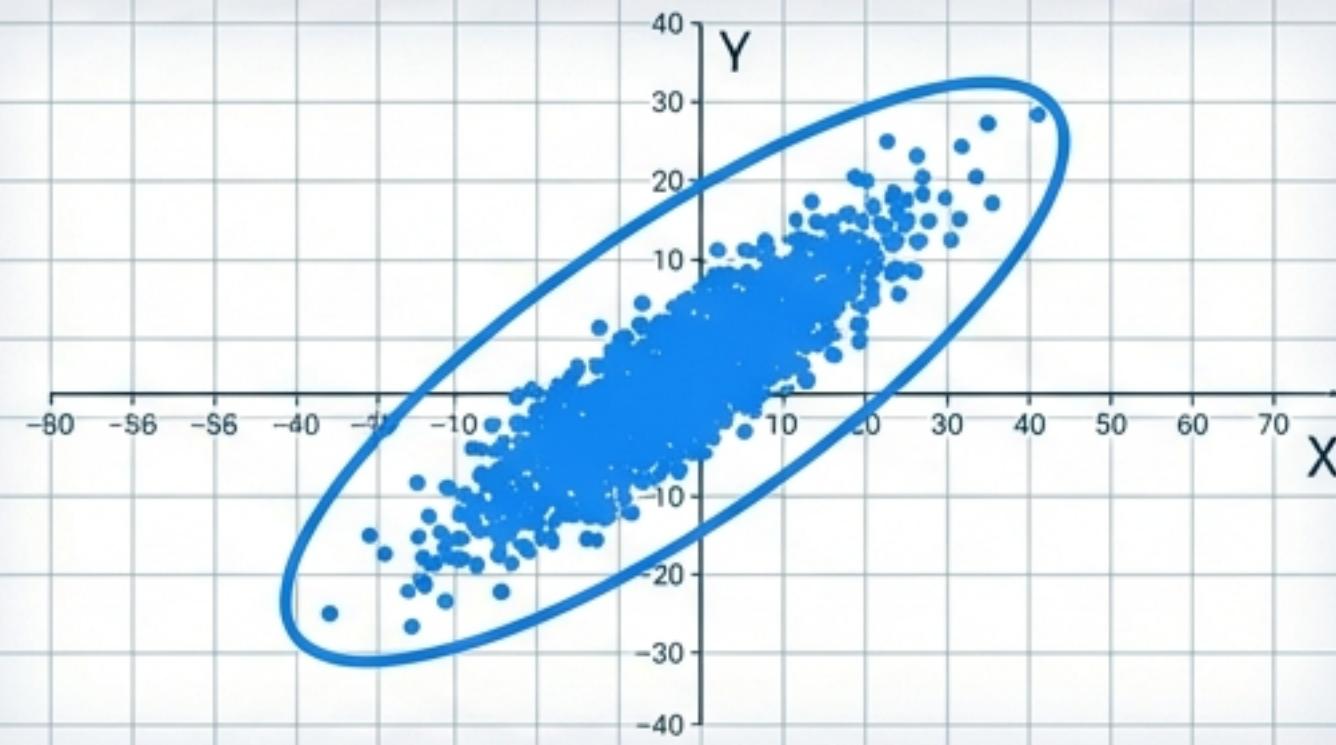
核心度量：馬氏距離 (The Core Metric: Mahalanobis Distance)

歐氏距離 (Euclidean Distance)



忽略相關性 (Ignores Correlation)

馬氏距離 (Mahalanobis Distance)



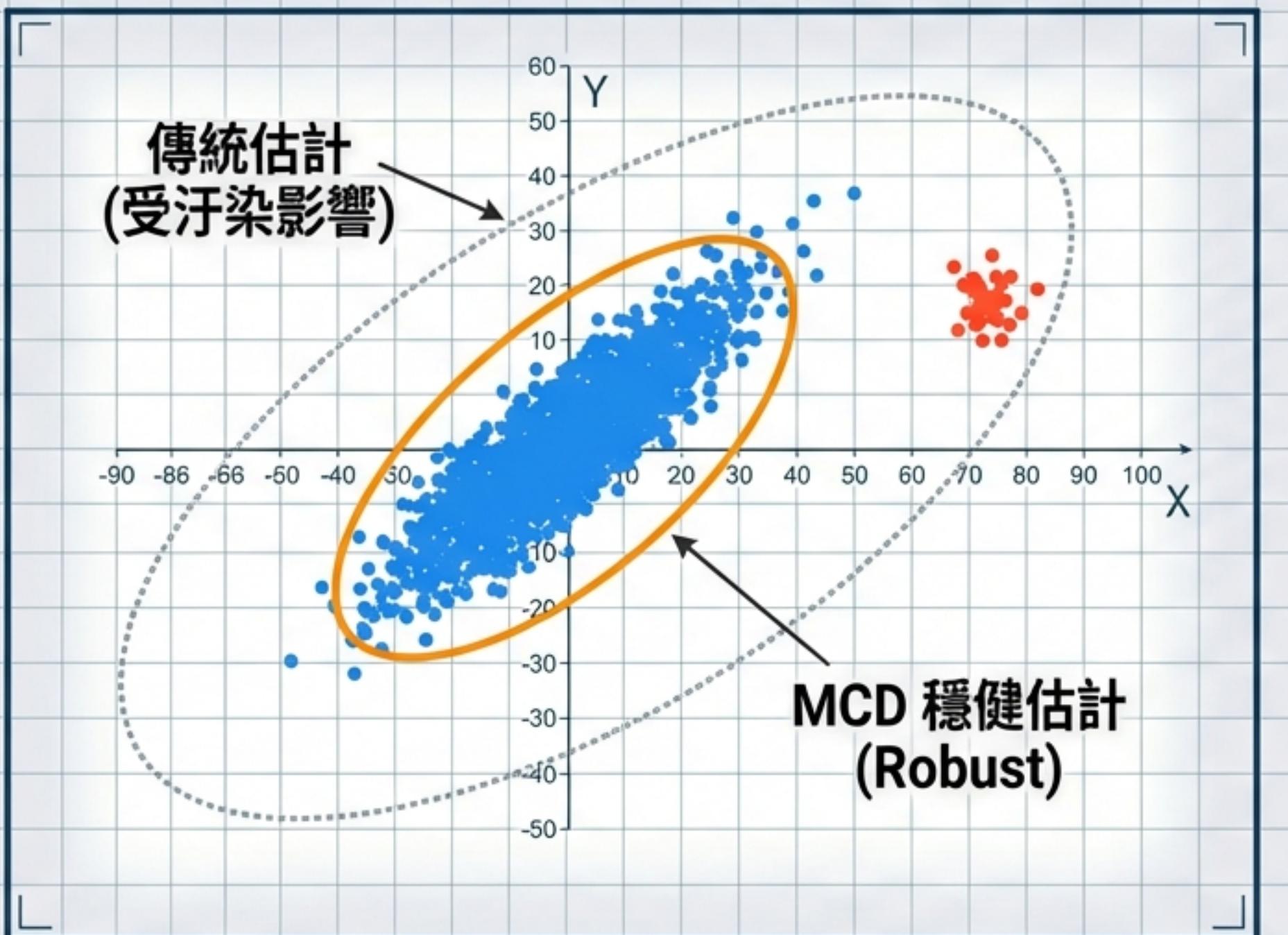
考慮相關性 (Accounts for Correlation)

$$D_M(x) = \sqrt{(x - \mu)^T \Sigma^{-1} (x - \mu)}$$

Σ^{-1} (Inverse Covariance) : 標準化變異數並消除變數間的相關性。

化工意義：當溫度 (T) 升高時，壓力 (P) 應同時升高。異常是「高溫低壓」。只有馬氏距離能檢測出這種違反物理相關性的異常。

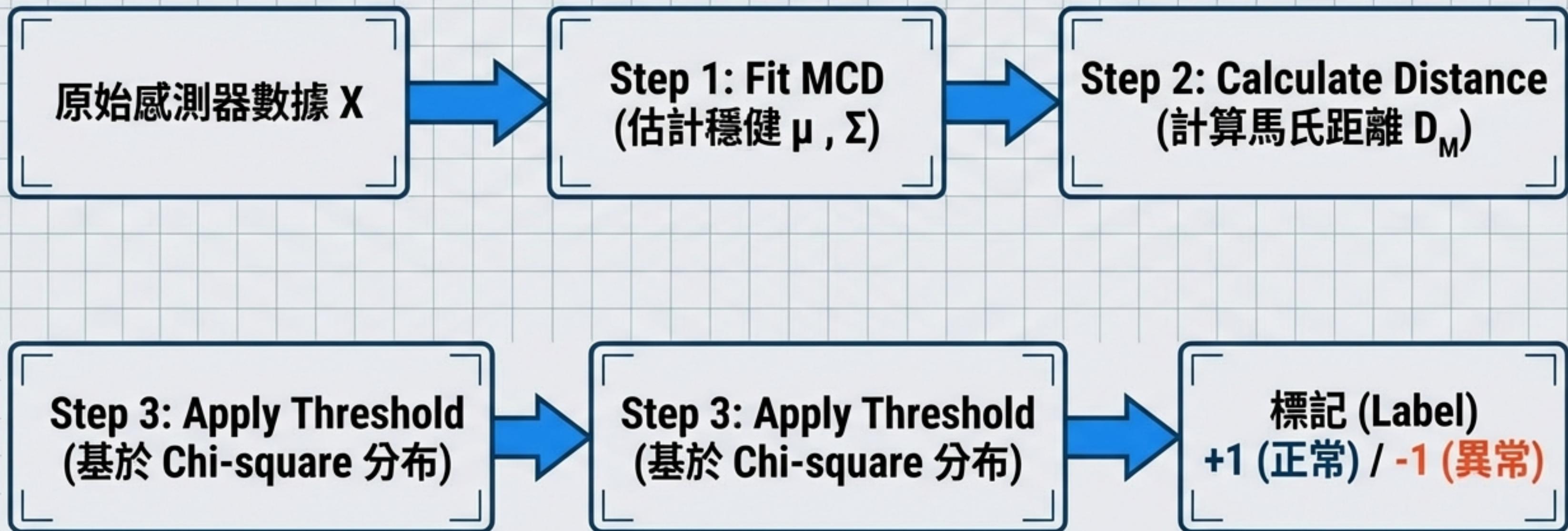
穩健性設計：MCD 估計器 (Robust Design: MCD Estimator)



概念：最小共變異數行列式
(Minimum Covariance Determinant, MCD)

- 問題：標準統計估計對離群值敏感。
- 方法：尋找數據中「最緊密」的子集 (h 個點)，使其共變異數行列式最小。
- 結果：自動排除訓練數據中的離群值，建立正確的「正常」模型。

演算法工作流程 (Algorithm Workflow)



閾值基於預期污染比例與自由度決定。

關鍵超參數設定 (Key Hyperparameters)

contamination (0.0 - 0.5)



- 定義：預期訓練數據中的異常比例。
- 影響：決定橢圓邊界的大小。
- 規則：值越大 = 邊界越緊 = 更多報警 (**More Alerts**)。

support_fraction (MCD Subset)



- 定義：MCD 演算法使用的子集比例。
- 影響：決定模型的穩健性 (**Robustness**)。
- 建議：預設值 (**None**) 通常已足夠，除非數據汙染極其嚴重。

程式實作 (Code Implementation)

```
from sklearn.covariance import EllipticEnvelope

# 1. 設定模型：預期 1% 的數據是異常的
model = EllipticEnvelope(contamination=0.01, random_state=42)

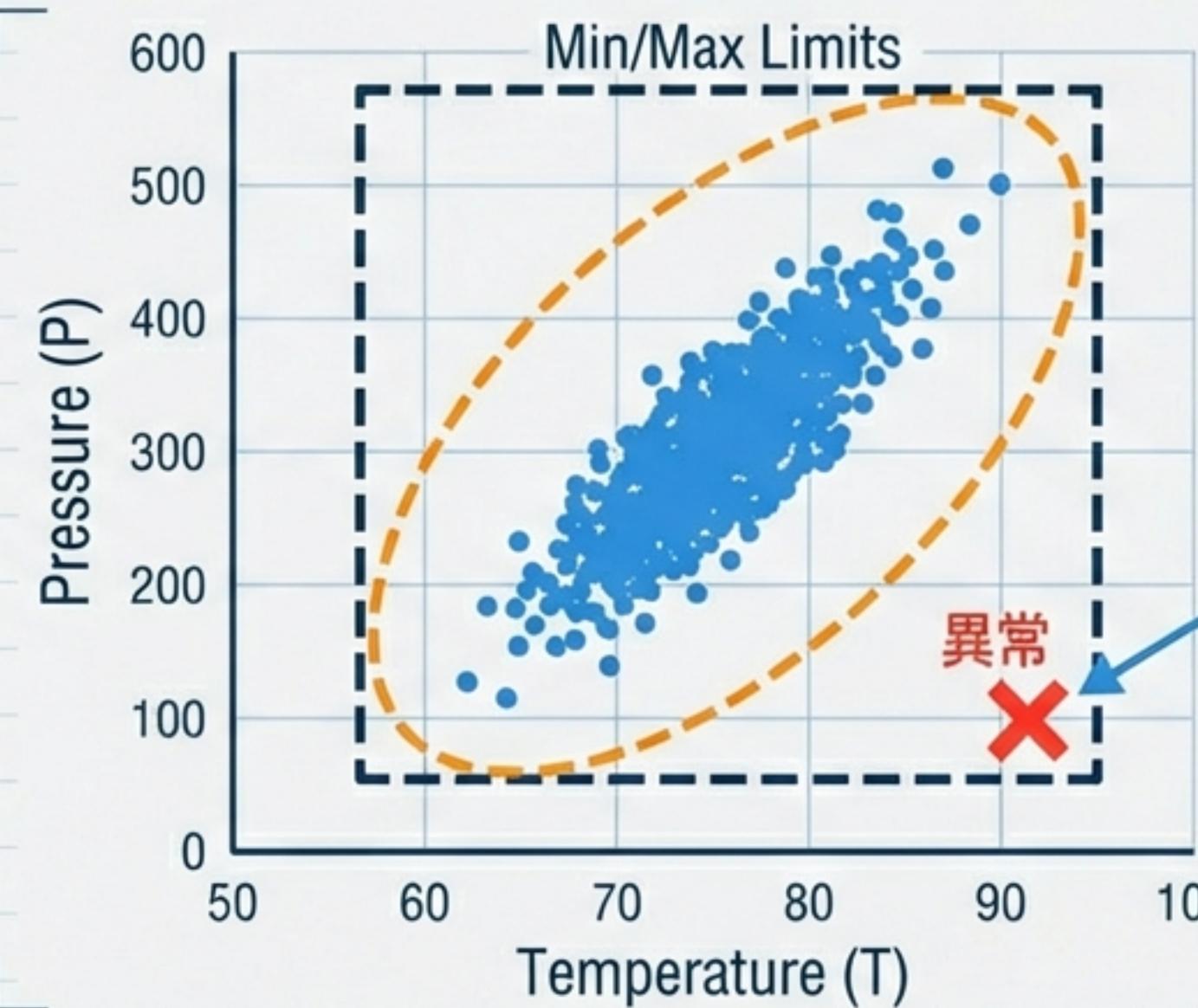
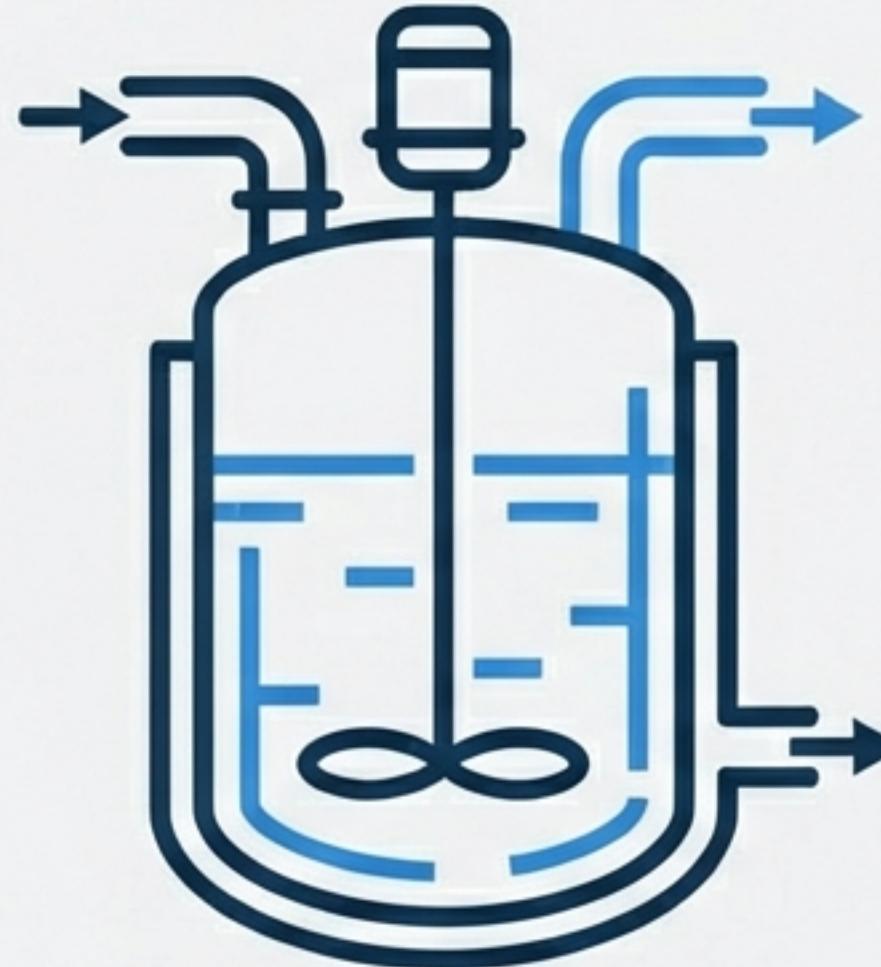
# 2. 訓練模型 (Fit)
model.fit(X_train)

# 3. 預測 (Predict): 1 = Normal, -1 = Anomaly
y_pred = model.predict(X_test)

# 4. 取得馬氏距離 (可用於監控嚴重程度)
dist = model.mahalanobis(X_test)
```

Pro Tip: 雖然馬氏距離能處理縮放，但為了數值穩定性，建議先進行標準化 (Standardization)。

實務應用：反應器品質監控 (Case Study: Reactor Quality Control)



場景：監控批次反應器的穩定狀態。

優勢：橢圓包絡能識別簡單高低限警報 (Hi/Lo Limits) 遺漏的異常，例如感測器漂移或效率下降。

演算法優缺點分析 (Pros & Cons)



優點 (Strengths)

$\pi \frac{1}{\div}$

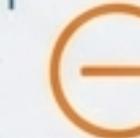
- 數學基礎 (Foundational)：
統計推論嚴謹，可解釋性強。



- 關聯性 (Correlation)：
完美處理化工變數間的耦合
(Coupling)。



- 穩健性 (Robustness)：
MCD 有效抵抗訓練數據污染。



限制 (Limitations)



- 高斯假設 (Gaussian Assumption)：數據必須接近常態分布。多峰分布效果差。

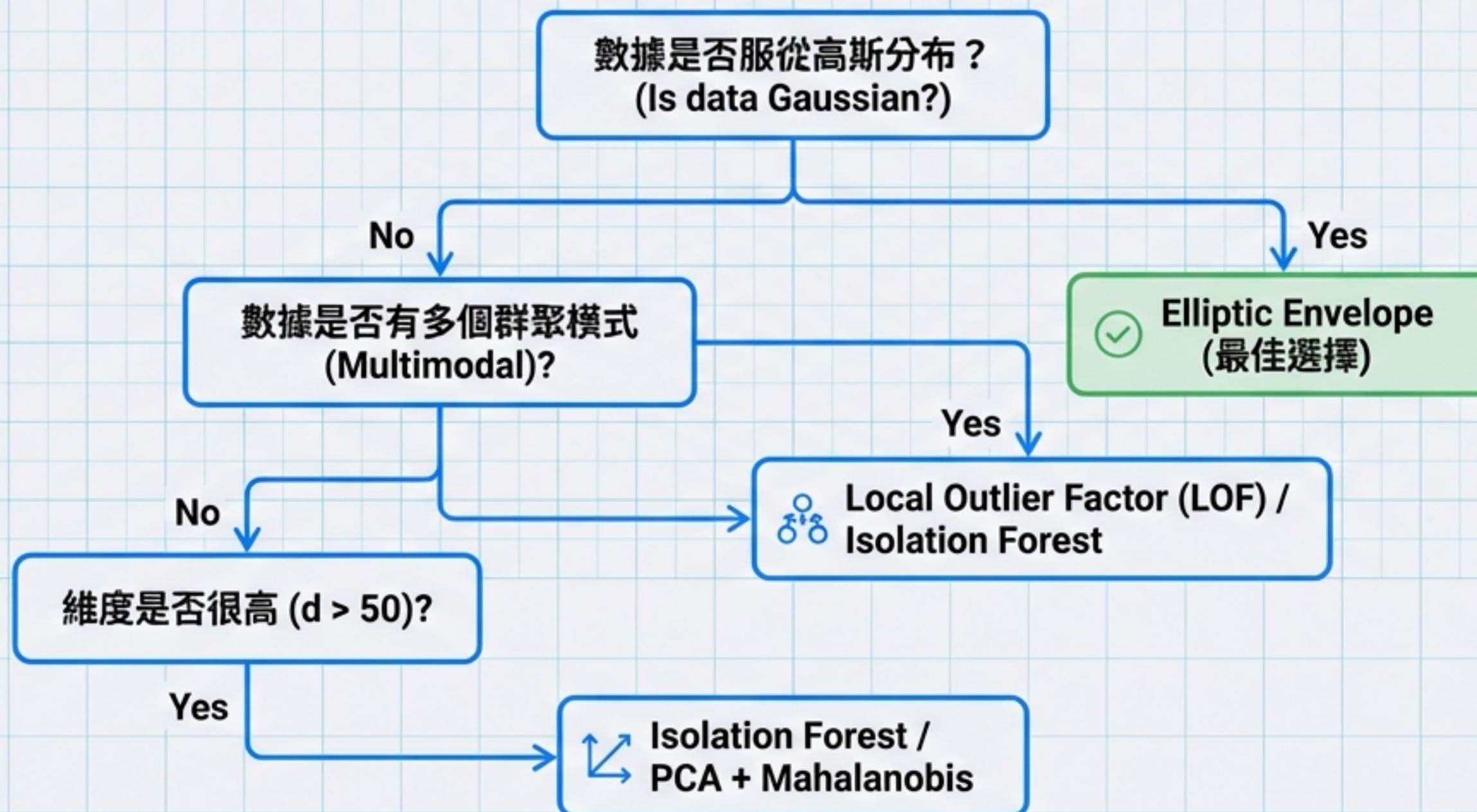


- 維度限制 (Dimensions)：
不適合極高維數據 ($d > 50$)。



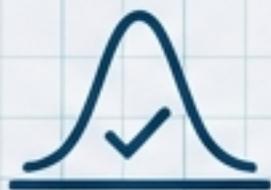
- 多操作模式 (Multimodal)：
無法同時處理啟動、關閉與穩態數據。

選擇指南：何時該用什麼？(Selection Guide)



Takeaway: 了解數據的「形狀」是選擇算法的第一步。

總結與下一步 (Summary & Next Steps)



高斯假設 (Gaussian)：適用於穩定的化工製程。



馬氏距離 (Mahalanobis)：解決變數間的相關性問題。



MCD 穩健估計：提供對髒數據的抵抗力。

下一步 (Next Unit) :

Unit 08 關聯規則學習 (Association Rule Learning)



「異常檢測不僅是發現錯誤，更是理解製程的正常邊界。」