

Unit 09 進階補充教材：製程安全異常偵測

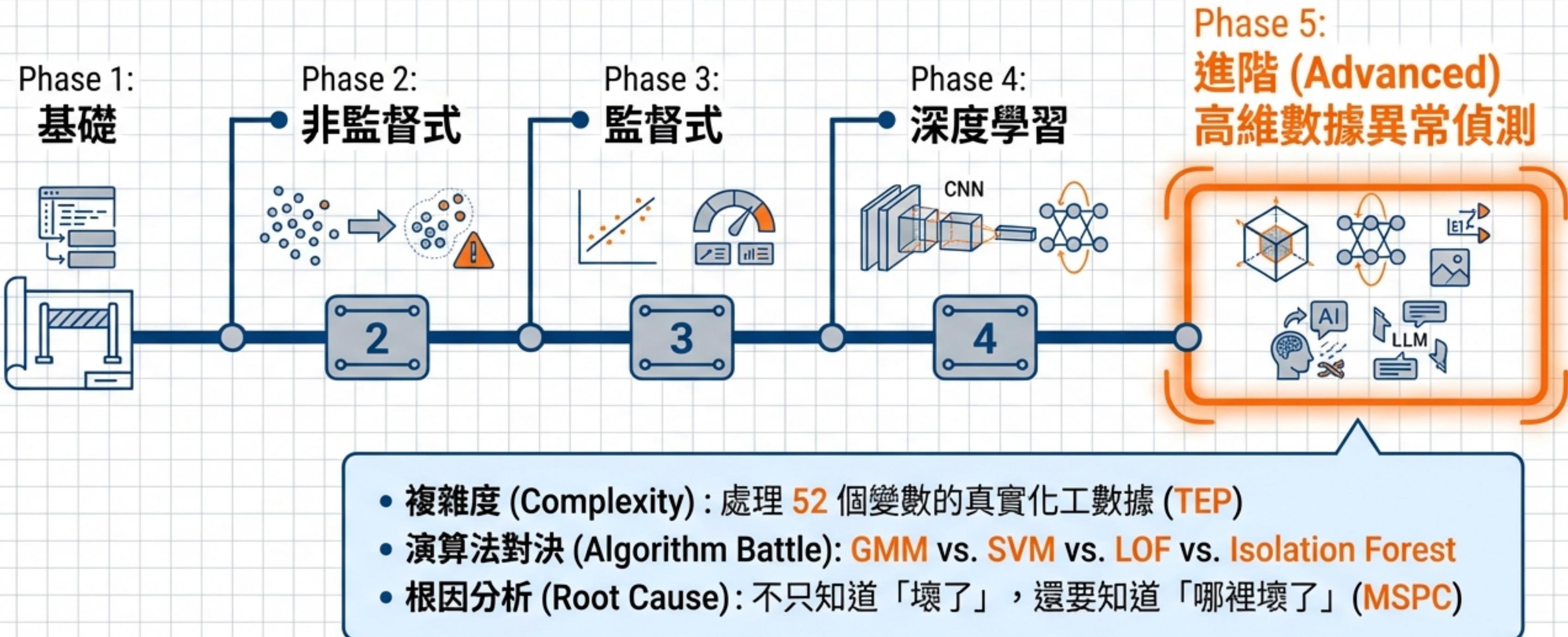
從 TEP 案例看高維數據的異常識別與根因分析

課程代號 CHE-AI-101 | 化工資料科學與機器學習實務

授課教師：莊曜禎 助理教授

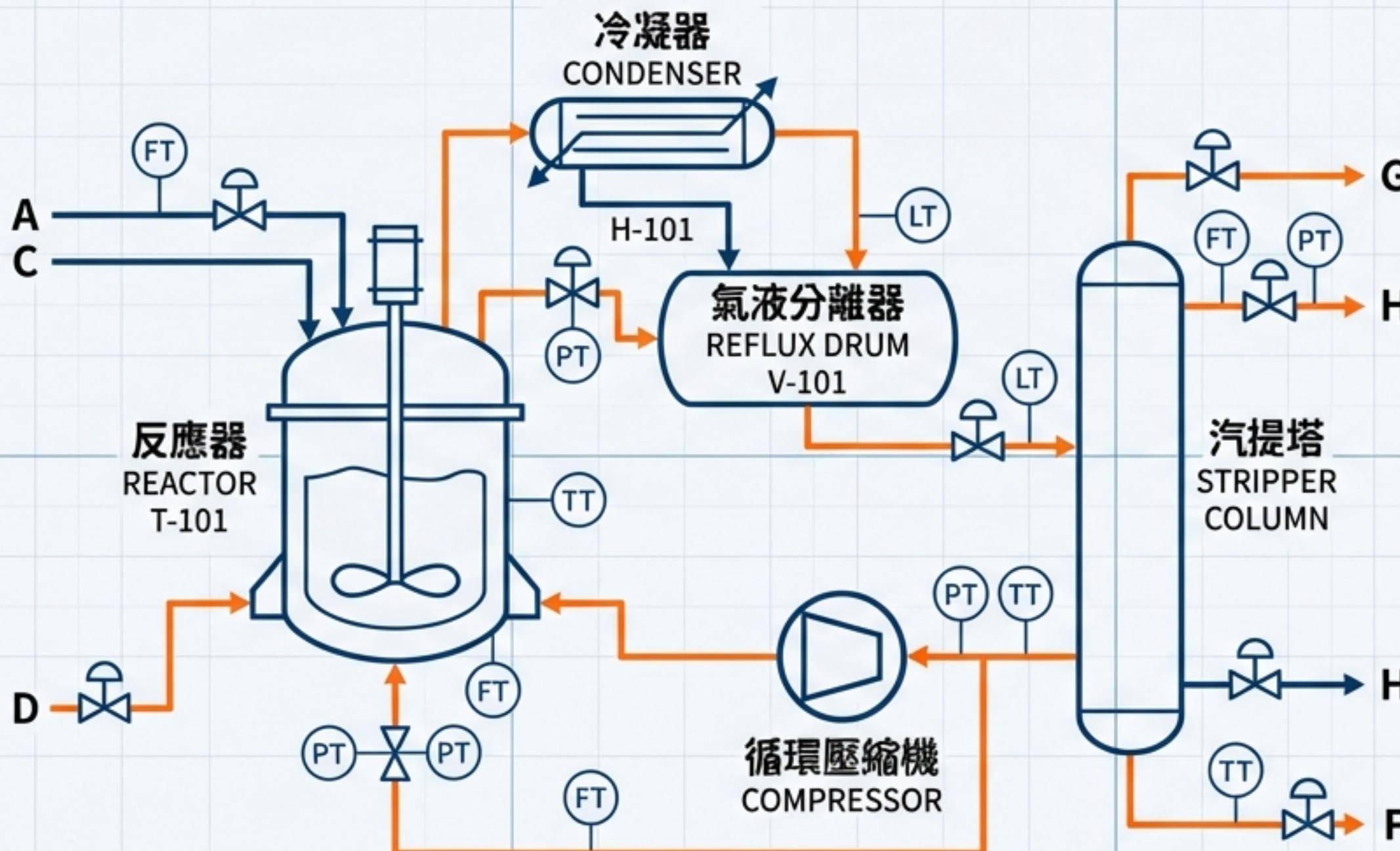
學期：114學年度第2學期

課程地圖：從基礎監控到進階診斷



■ 適用對象：已掌握 Unit 01-08 內容，欲深入理解數學原理與實務部署的學員

挑戰場景：Tennessee Eastman Process (TEP)



數據規模 (Data Scale)

- 變數 (Variables): 52 個 (41 量測 + 11 操作)
- 反應 (Reactions): 4 反應物 (A,C,D,E) → 2 產品 (G,H) + 副產物 F
- 故障 (Faults): IDV1 - IDV20 (含 Step, Random, Drift, Sticking)
- 取樣頻率: 每 3 分鐘一筆數據

Key Insight: TEP 是過程控制領域的 MNIST，但具備更強的物理交互作用與動態特性。

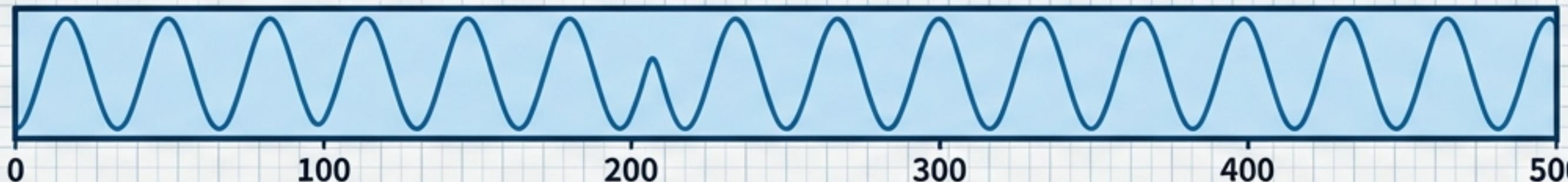
理解製程動態：真實數據的多時間尺度特性

XMEAS_07 反應器壓力 (Slow Drift)



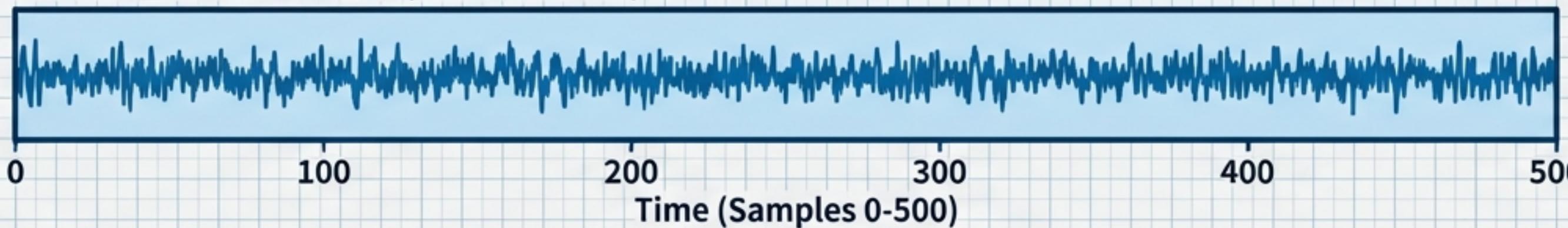
$\tau_{\text{slow}} \approx 5 \text{ hr}$
(整體製程漂移)

XMEAS_08 反應器液位 (Oscillation)



$\tau_{\text{fast}} \approx 10 \text{ min}$
(PID 控制器振盪)

XMEAS_09 反應器溫度 (White Noise)



$\tau_{\text{medium}} \approx 1 \text{ hr}$
(測量噪聲/反應動力學)

Takeaway: 真實數據不是 i.i.d. (獨立同分布)，它具有慣性與記憶 (Inertia and Memory)。

PCA 深度解析：高維數據的長尾分佈

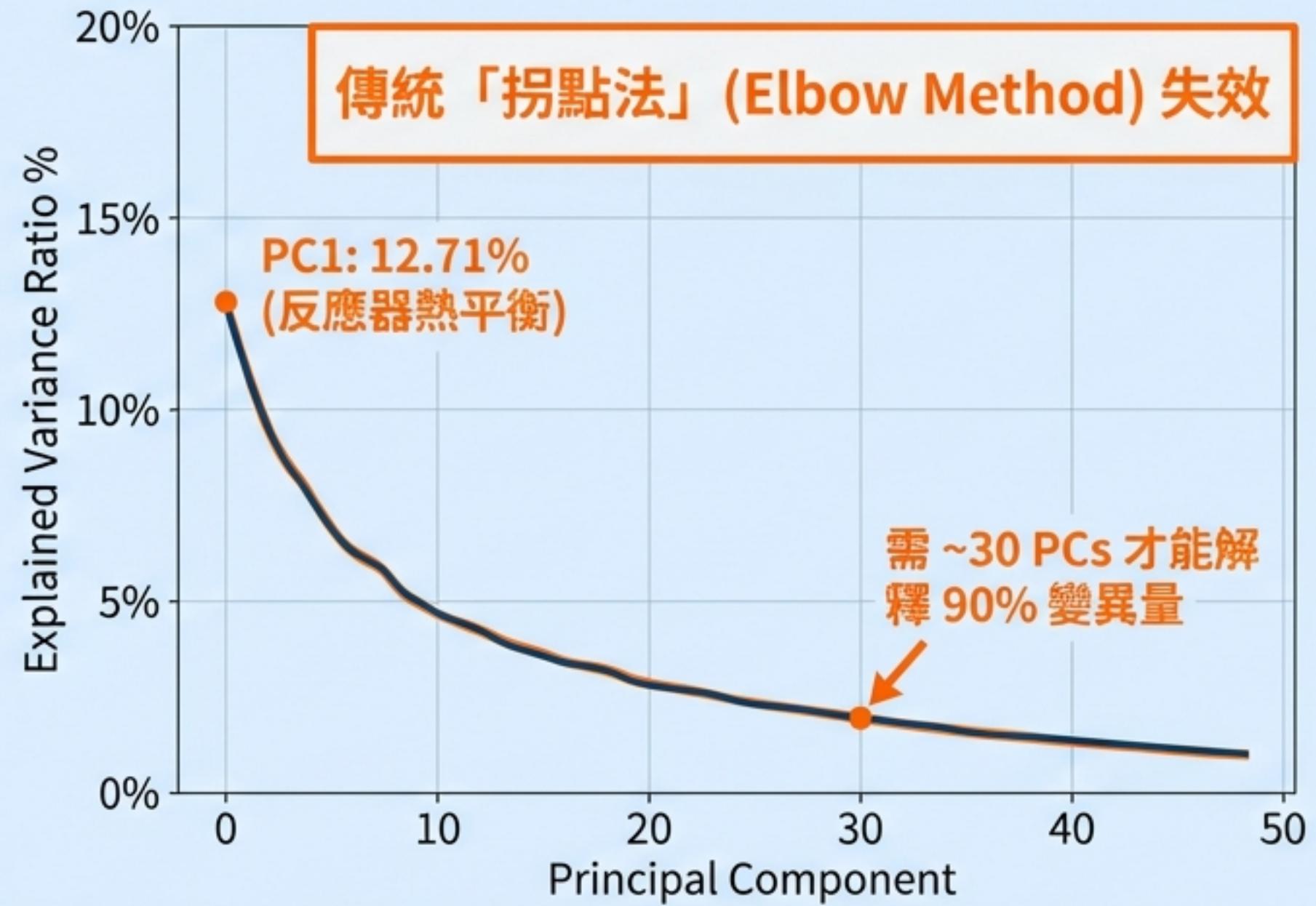
核心原理 (Core Math)

$$z = \frac{x - \mu}{\sigma}$$

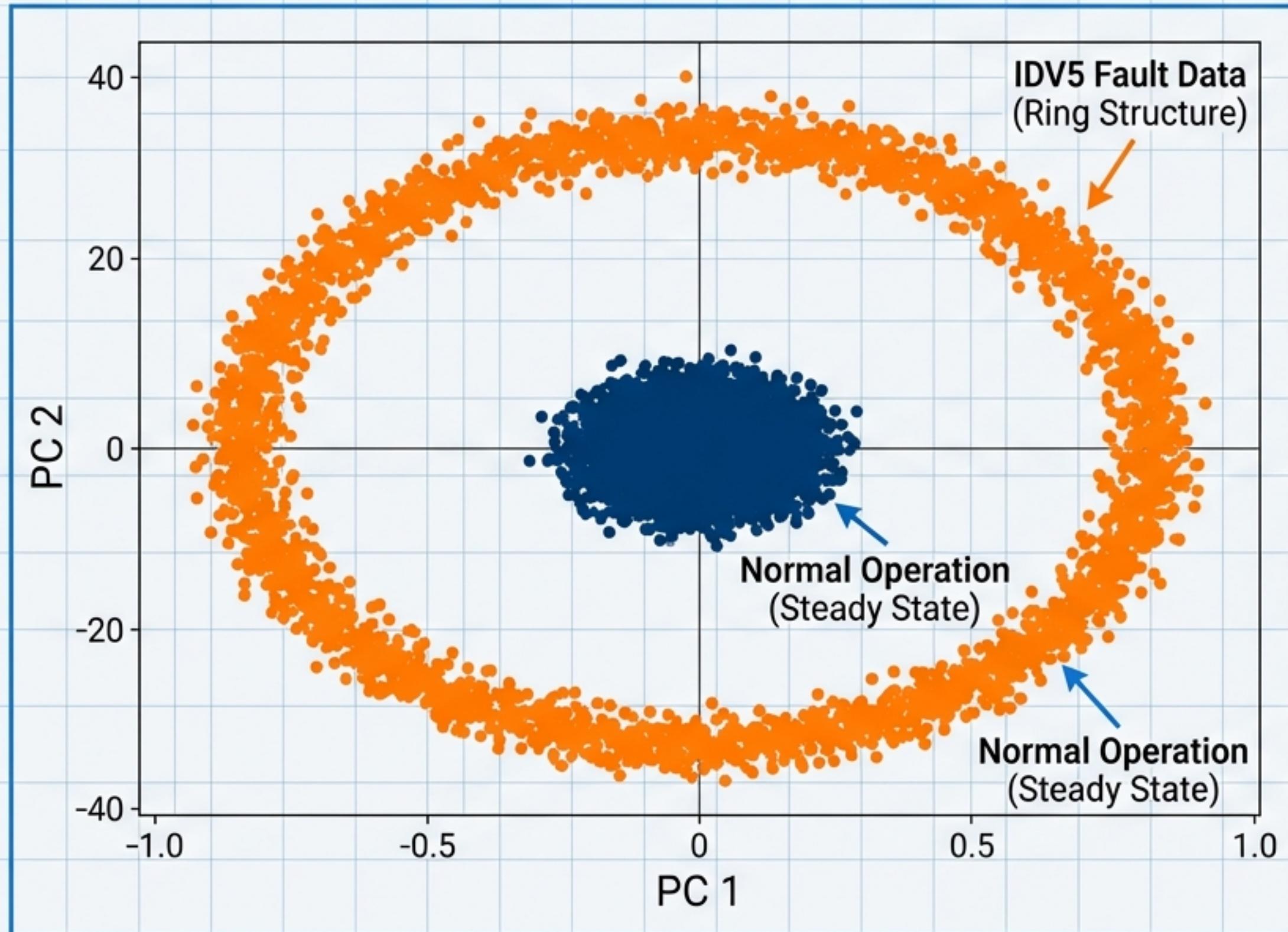
$$\sum v = \lambda v$$

與簡單數據集不同，TEP
數據的資訊高度分散。

Scree Plot (碎石圖)



看見異常：IDV5 (冷凝器冷卻水故障) 視覺化

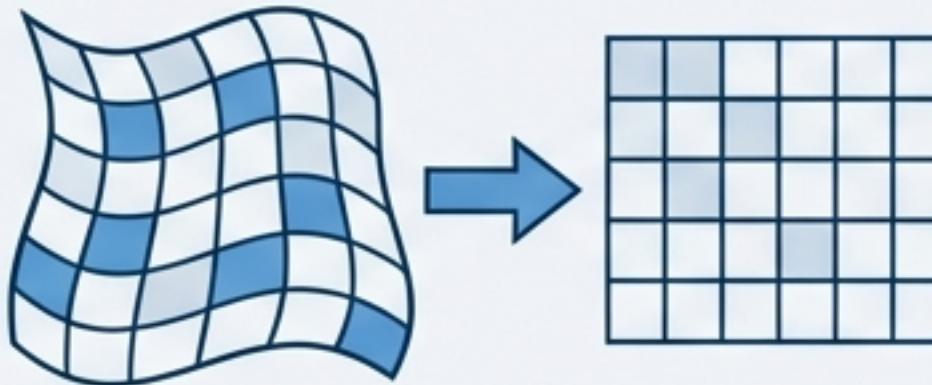


故障特徵解析

- 故障類型: 冷凝器冷卻水入口溫度階躍變化 (Step Change)
- 物理機制: 系統控制器試圖自我調節，導致數據在 PCA 空間形成軌道，但無法回到中心。
- Mahalanobis Distance:
 $D_M^2 > 50$ (Fault) vs < 9.21 (Normal)

超越標準 PCA：進階變種的成與敗

Kernel PCA (RBF)

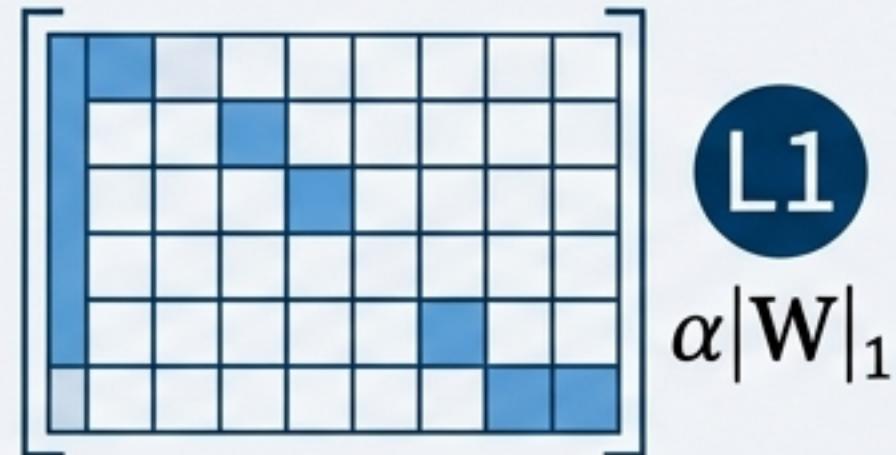


- 原理：映射到高維特徵空間
處理非線性

✗ 不推薦用於 IDV5

過度平滑 (Over-smoothing)
破壞了線性的環形結構，導致正常與故障數據重疊。

Sparse PCA (L1)

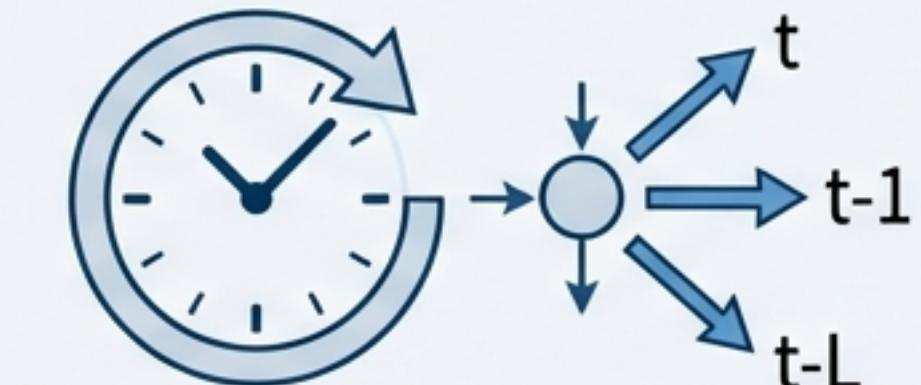


- 原理：引入 L1 罰項 ($\alpha|W|_1$)
強制負荷矩陣稀疏化

✓ 極佳的可解釋性

將 PC1 相關變數從 52 個減至 27 個，明確指出「反應器變數組」主導變異。

Dynamic PCA

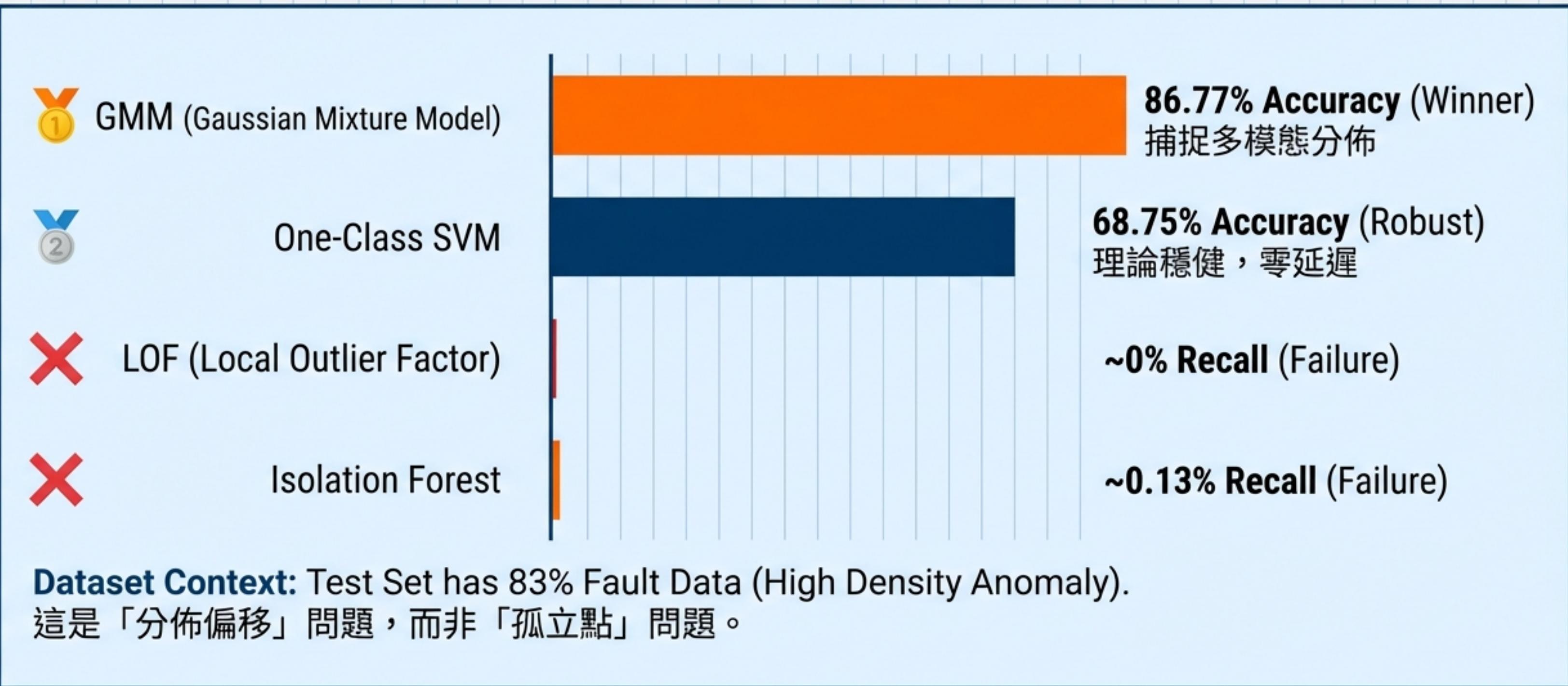


- 原理：引入時間延遲變數
(t, t - 1, t - L)

I 捕捉動態

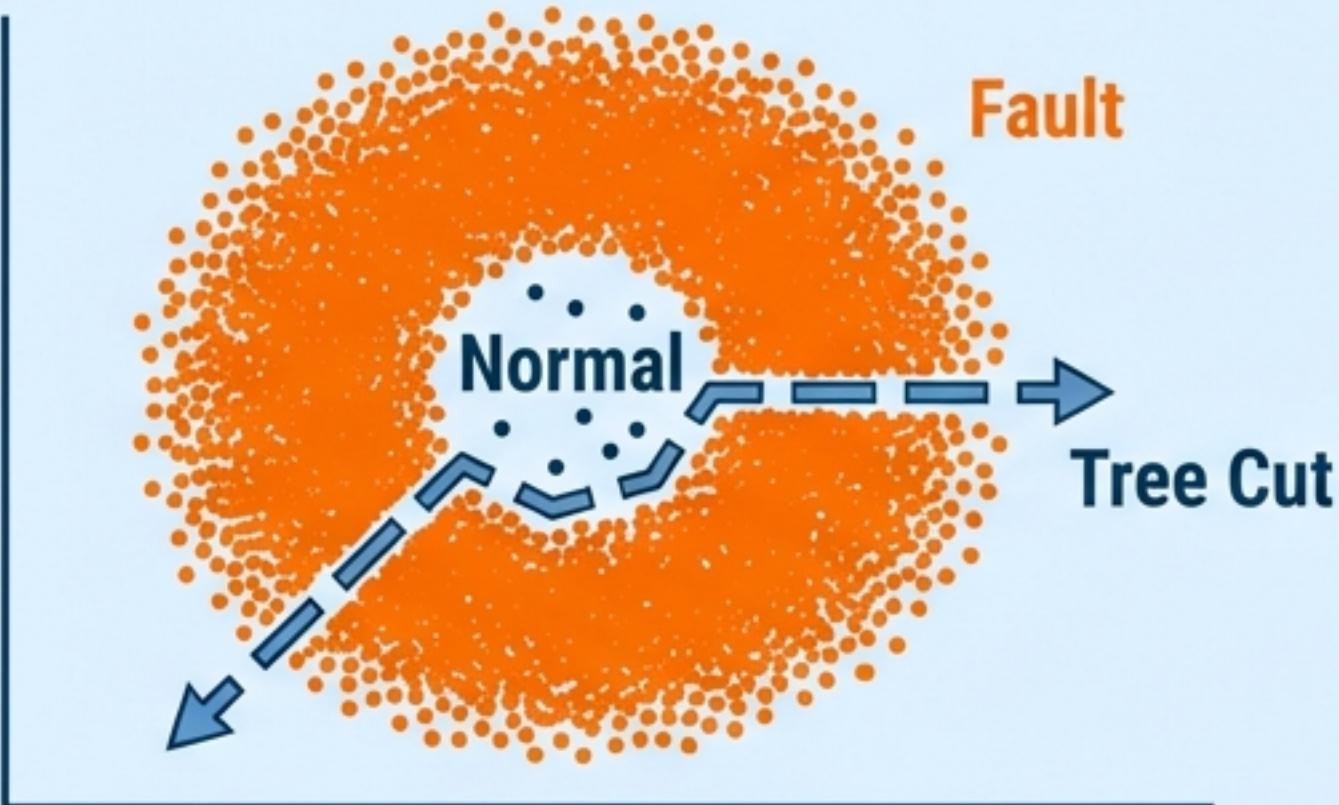
用於捕捉 PID 控制滯後與反應器動態。

演算法大對決：TEP IDV5 異常偵測效能



為何傳統方法失效？假設與現實的落差

Isolation Forest 失效原因

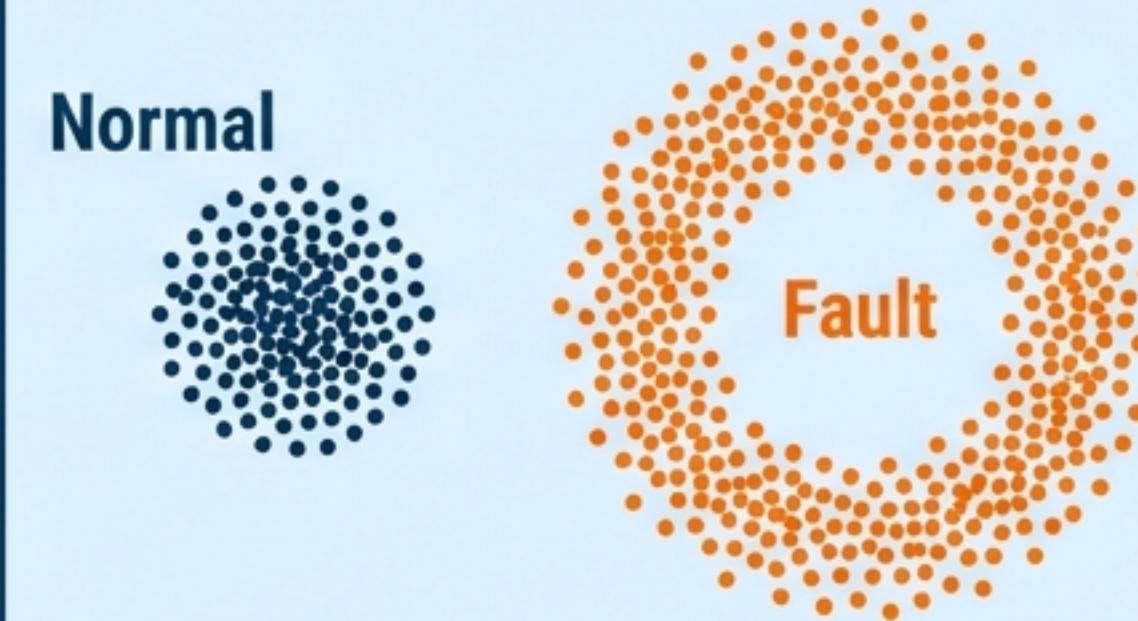


假設：異常是稀疏且孤立的。

現實：故障樣本佔 83% (密集區)。

結果：樹模型認為故障區才是「正常」，隔離了真正的正常點！

LOF 失效原因



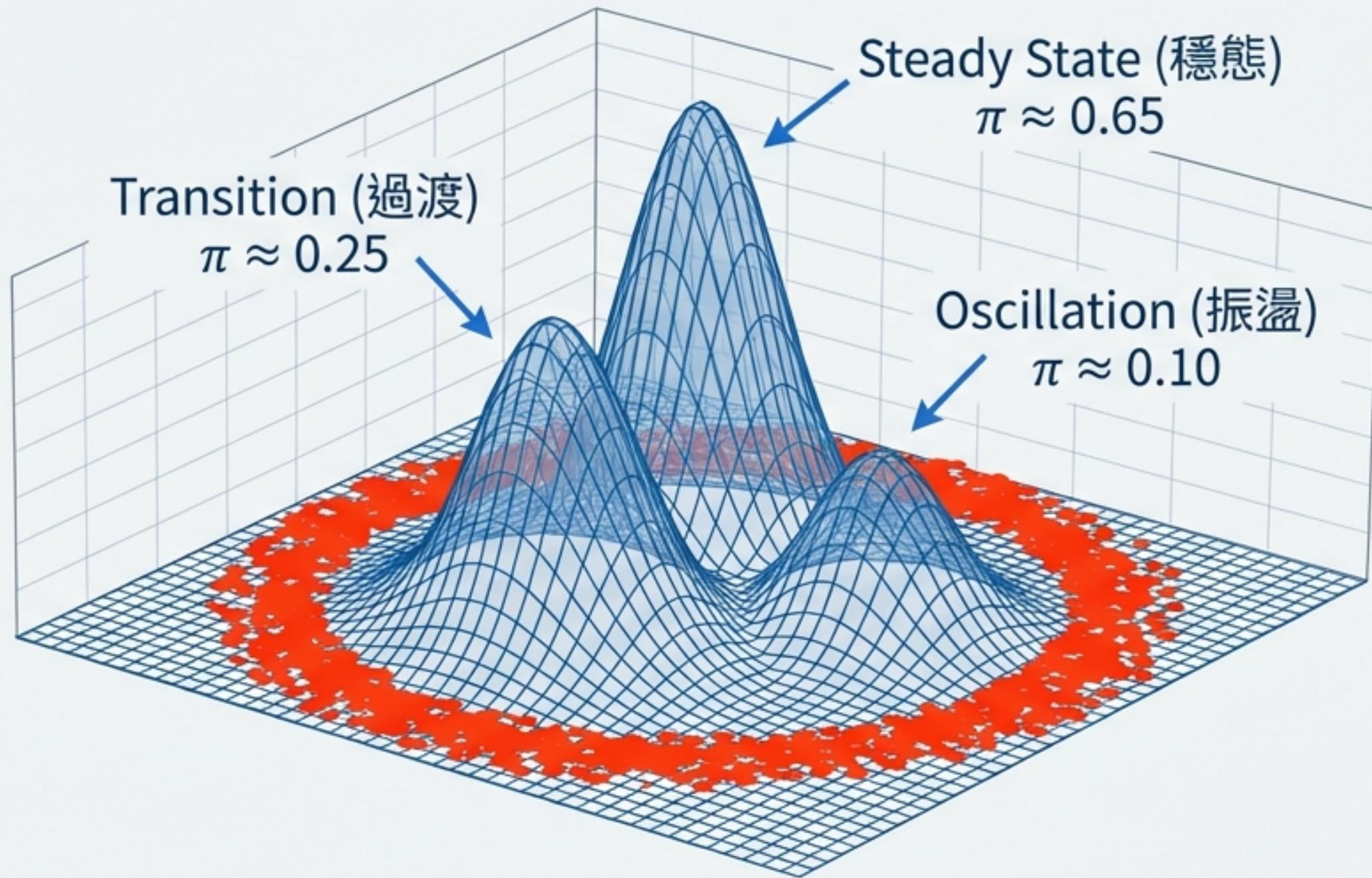
假設：異常點的局部密度 < 鄰居密度。

現實：故障的「環形結構」密度 ≈ 正常「核心」密度。

結果：LOF 分數 ≈ 1，無法區分。

Critical Lesson: Understand your data topology before choosing your model.

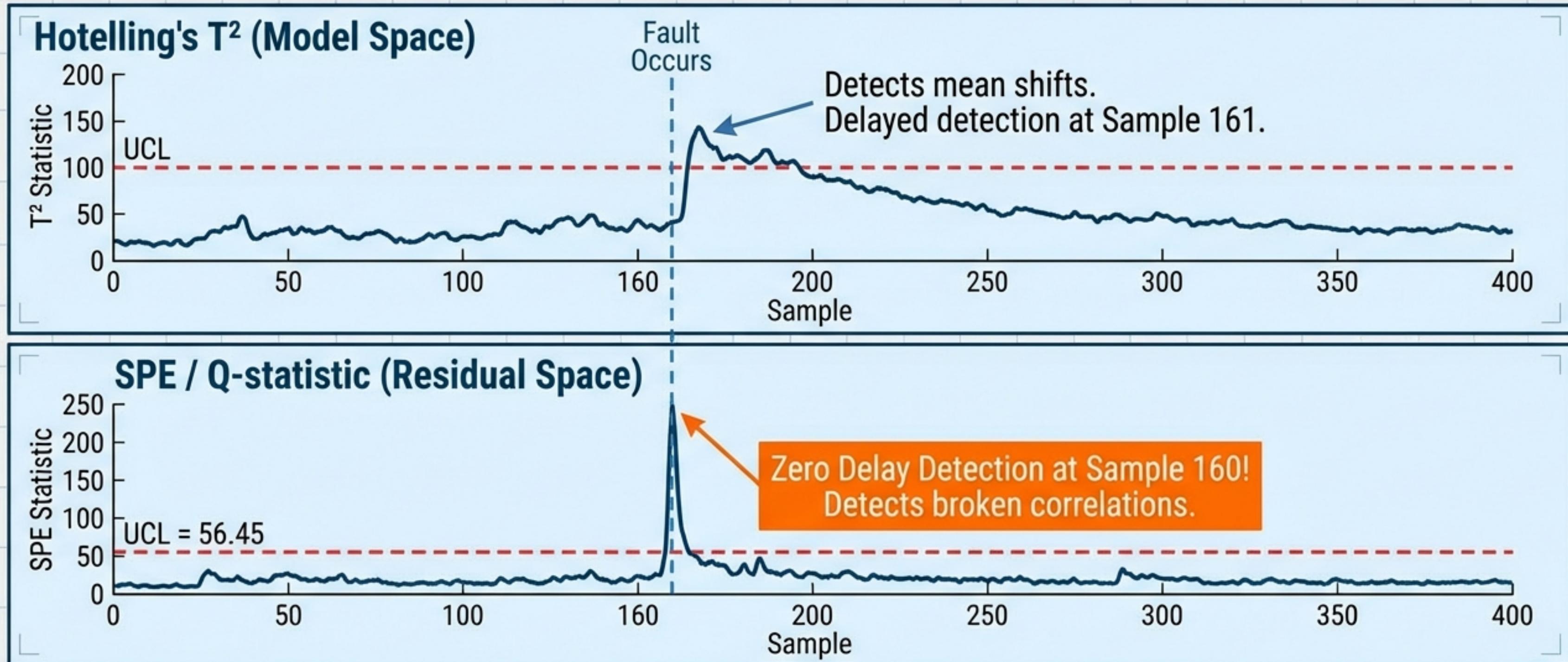
冠軍解析：高斯混合模型 (GMM)



$$\text{Math: } p(x) = \sum \pi_k \mathcal{N}(x | \mu, \Sigma)$$

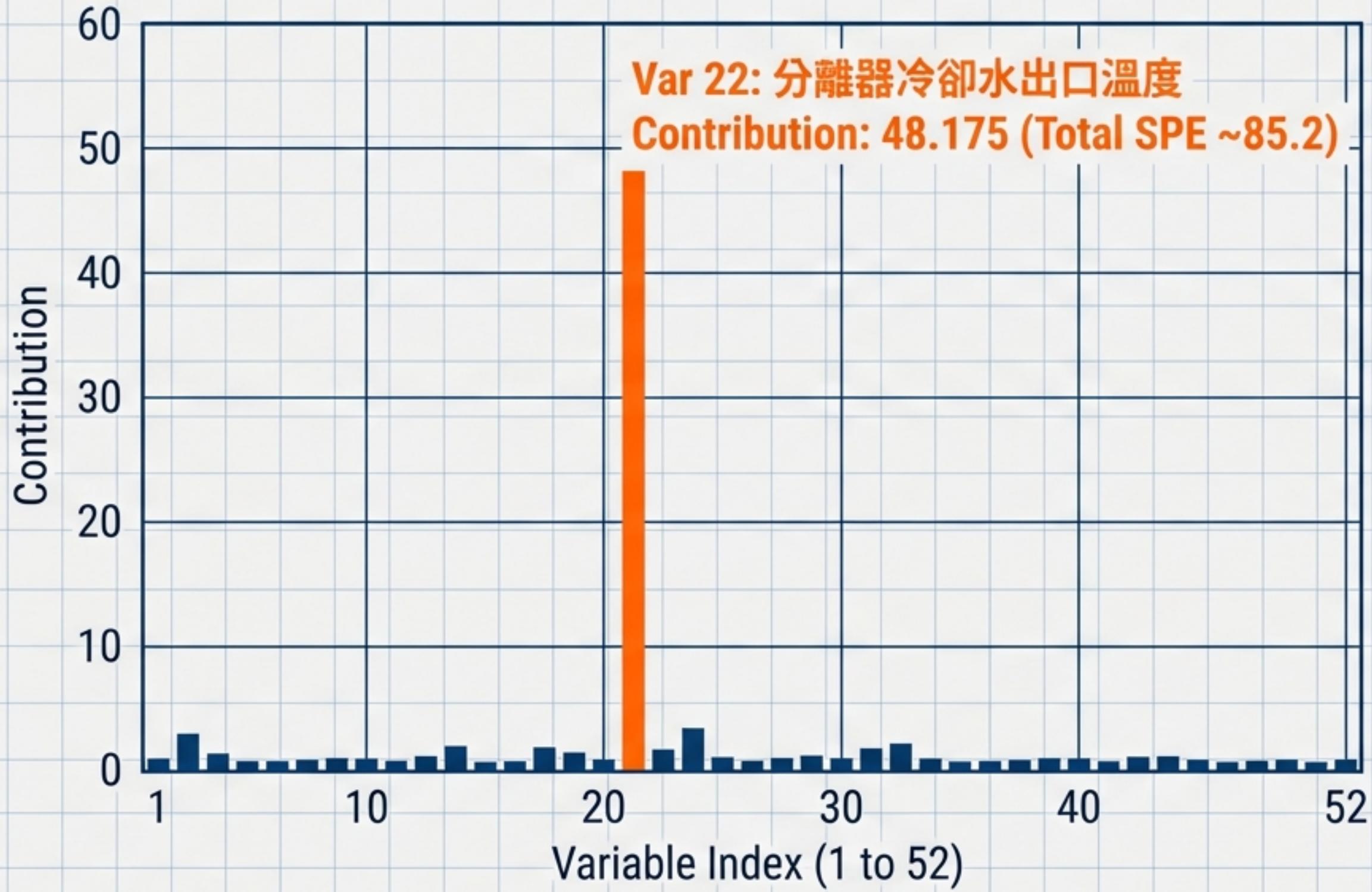
Why it Won: GMM 建立了「正常」的結構模型 (Structure of Normality)。任何落在這三座山之外的數據都被準確判定為異常。

MSPC 診斷： T^2 與 SPE 的互補性



Key Insight: IDV5 在改變變數平均值之前，先破壞了變數間的相關結構 (Correlation Structure)，因此 SPE 更敏感。

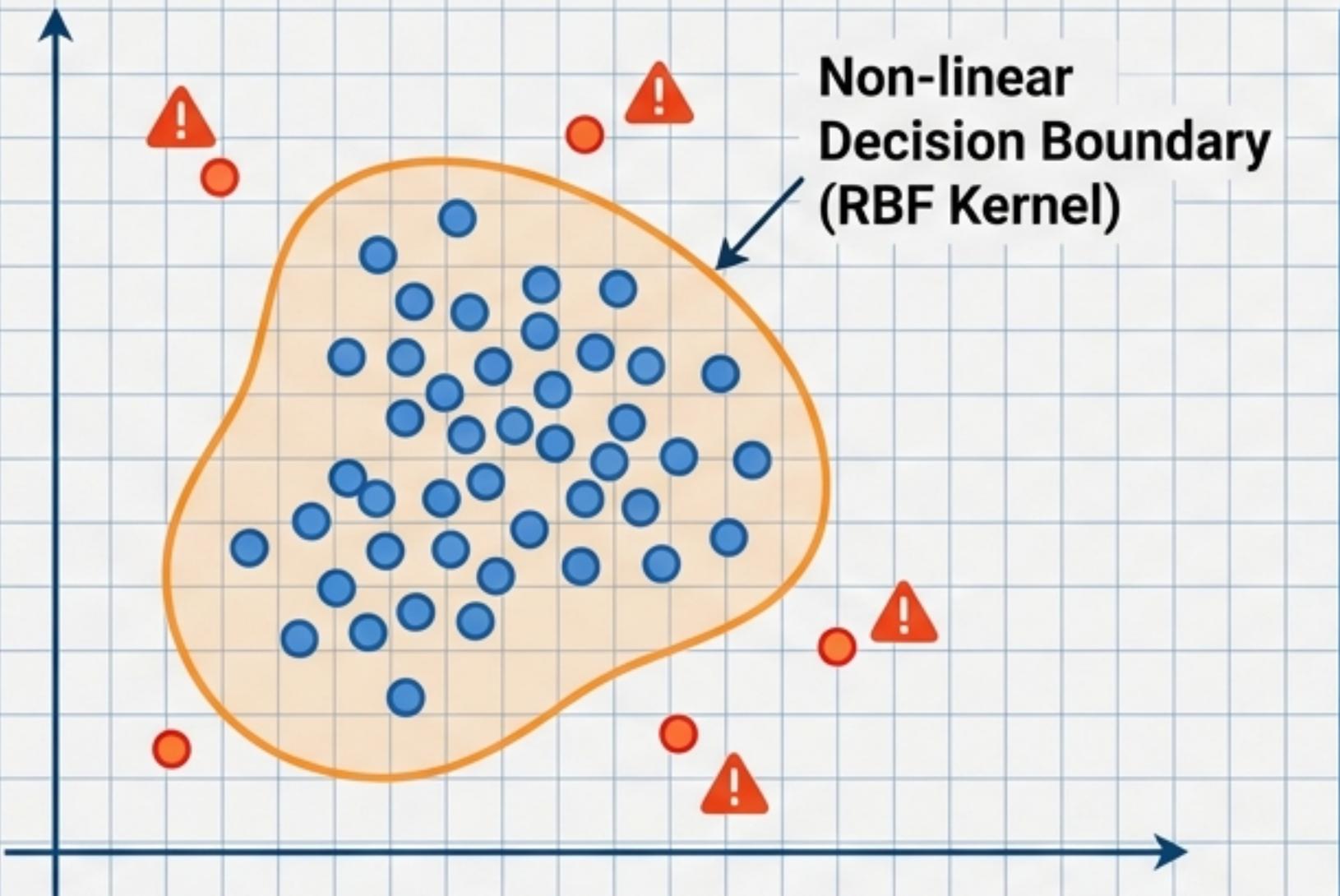
根因分析：誰觸發了警報？



診斷報告 (Diagnosis)

- 異常源頭: 冷卻系統 (Cooling System)
- 行動建議: 立即檢查冷凝器/分離器冷卻迴路 (Condenser/Separator cooling loop)
- 價值: AI 將「黑盒子」轉變為可操作的「診斷工具」。

穩健之選：One-Class SVM



Technical Specs Panel.

- **Kernel:** RBF (Radial Basis Function) - 表現最佳 (93.7% detection with tuning)
- **Parameter ν (Nu):** 0.04 (設定預期異常比例)
- **Mechanism:** 在高維特徵空間尋找包圍正常數據的超球面 (Hypersphere)

Verdict: 工業現場的首選之一。無須假設數據分佈 (Distribution-free)，且凸優化 (Convex Optimization) 保證全域最優解。

實務部署策略：分層監控體系

Phase 1: 快速原型 (Prototype)

Tool: **Isolation Forest**

Note: Fast, no tuning needed, baseline check.



Maintenance: Quarterly retraining to handle **Concept Drift** (e.g., catalyst aging).

Phase 2: 生產部署 (Production)

Tool: **GMM + MSPC (SPE)**

Note: **GMM** for high accuracy detection; **SPE** for root cause diagnosis.

Phase 3: 極致優化 (Optimization)

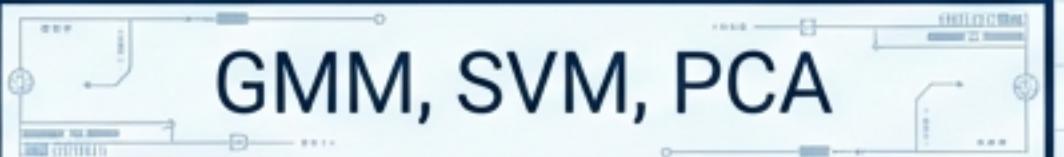
Tool: **Ensemble Voting (GMM + SVM)**

Note: Reduce false positives.

結語：定義未來的工程師

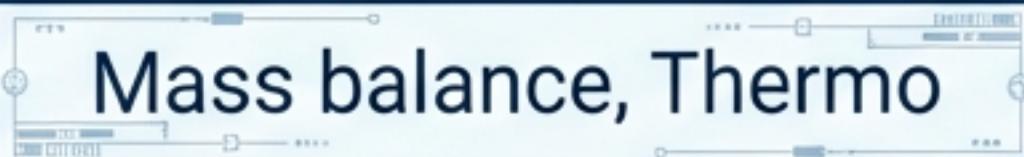


Algorithms are Tools



GMM, SVM, PCA

Physics is the Guide



Mass balance, Thermo

Data is the Fuel



TEP, SCADA logs

“AI 不會取代化工工程師；但懂得使用 AI 的化工工程師將取代不懂的人。”

Next Step: 下載 TEP 數據集，運行 Jupyter Notebook，親自驗證這些演算法。