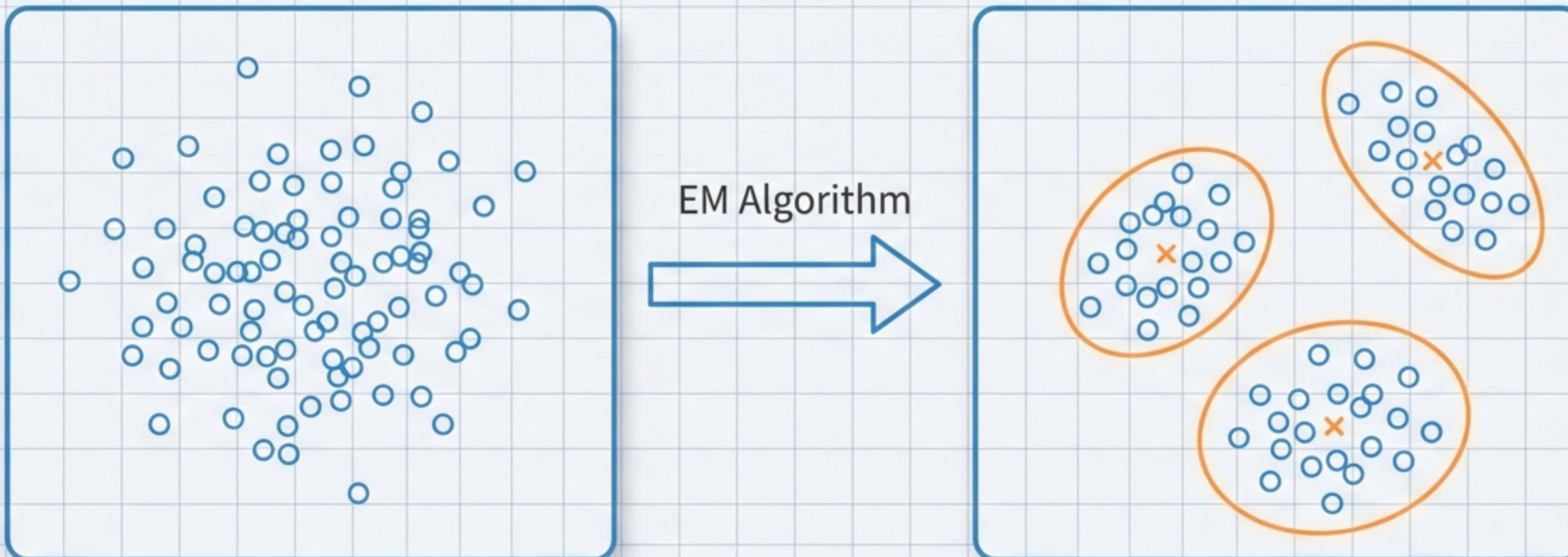


Unit 05: 高斯混合模型 (Gaussian Mixture Models)

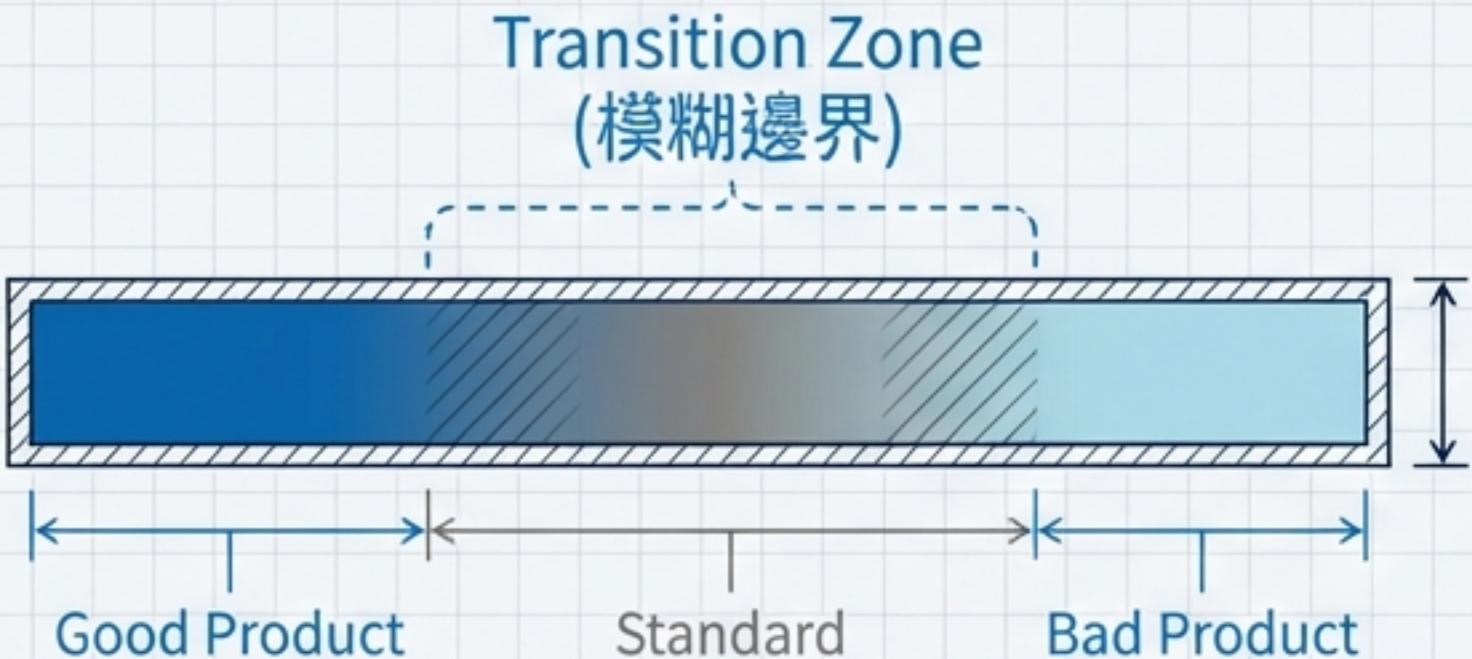
化工製程中的機率分群與不確定性量化 (Probabilistic Clustering & Uncertainty Quantification)



挑戰：當「硬分群」遇上真實世界的模糊邊界

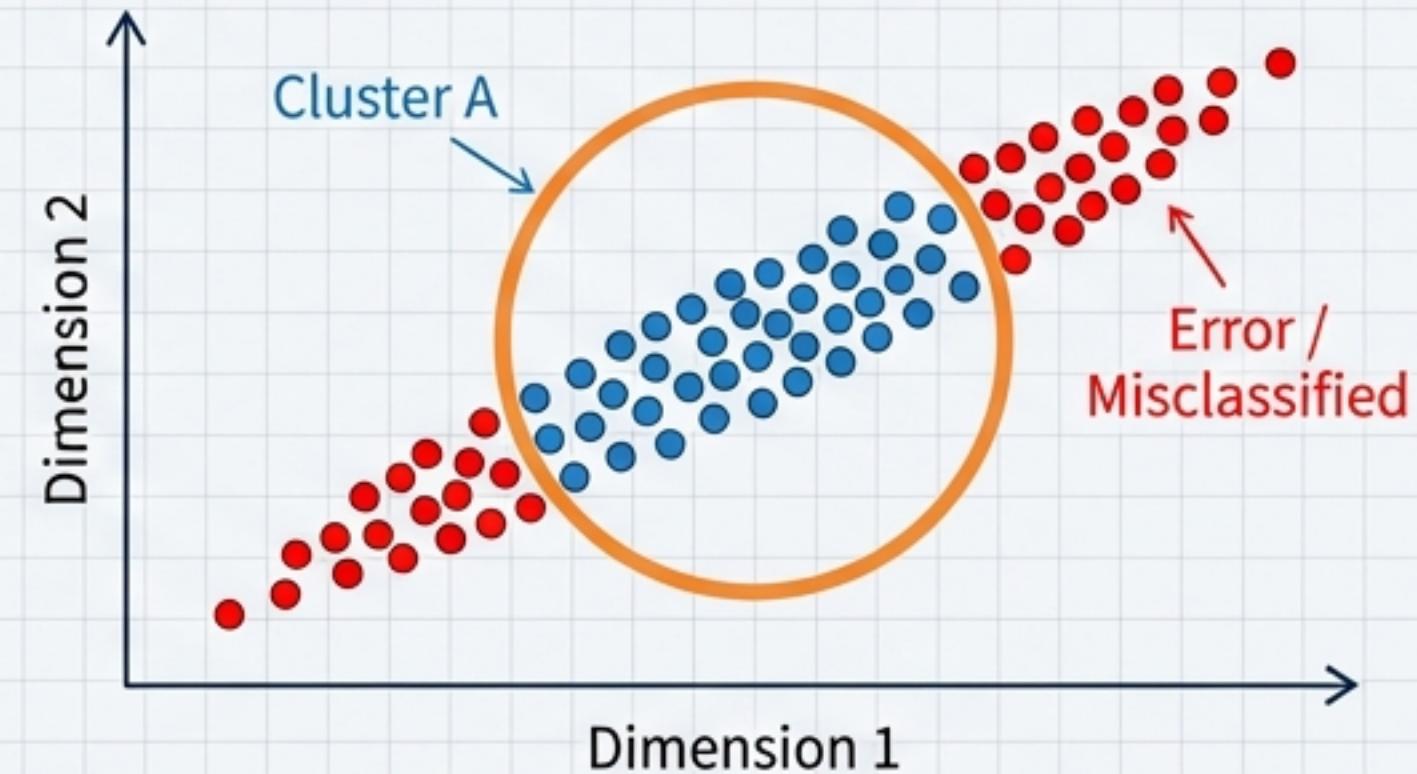
化工現狀 (Engineering Reality)

產品品質通常呈現連續的梯度分布，而非非黑即白的二元分類。



K-Means 的局限 (Limitation)

- 強制硬分群 (Hard Clustering)：非 A 即 B
- 假設球形分布 (Spherical Assumption)
- 在重疊區域容易產生誤判

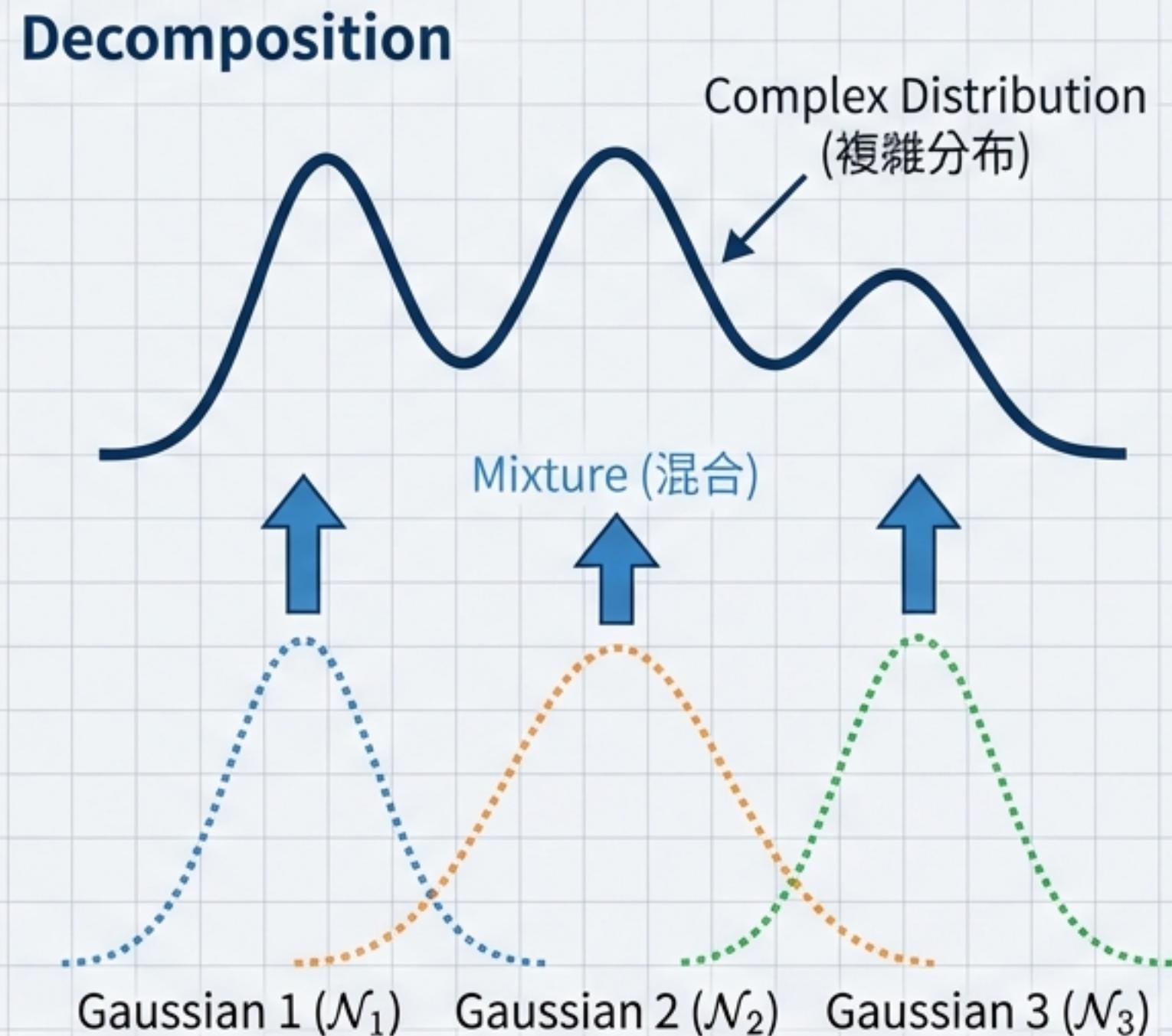


Key Insight: 大自然並不總是畫出完美的圓。我們需要一個能理解重疊與機率的模型。

GMM 核心原理：基於機率的軟分群視角

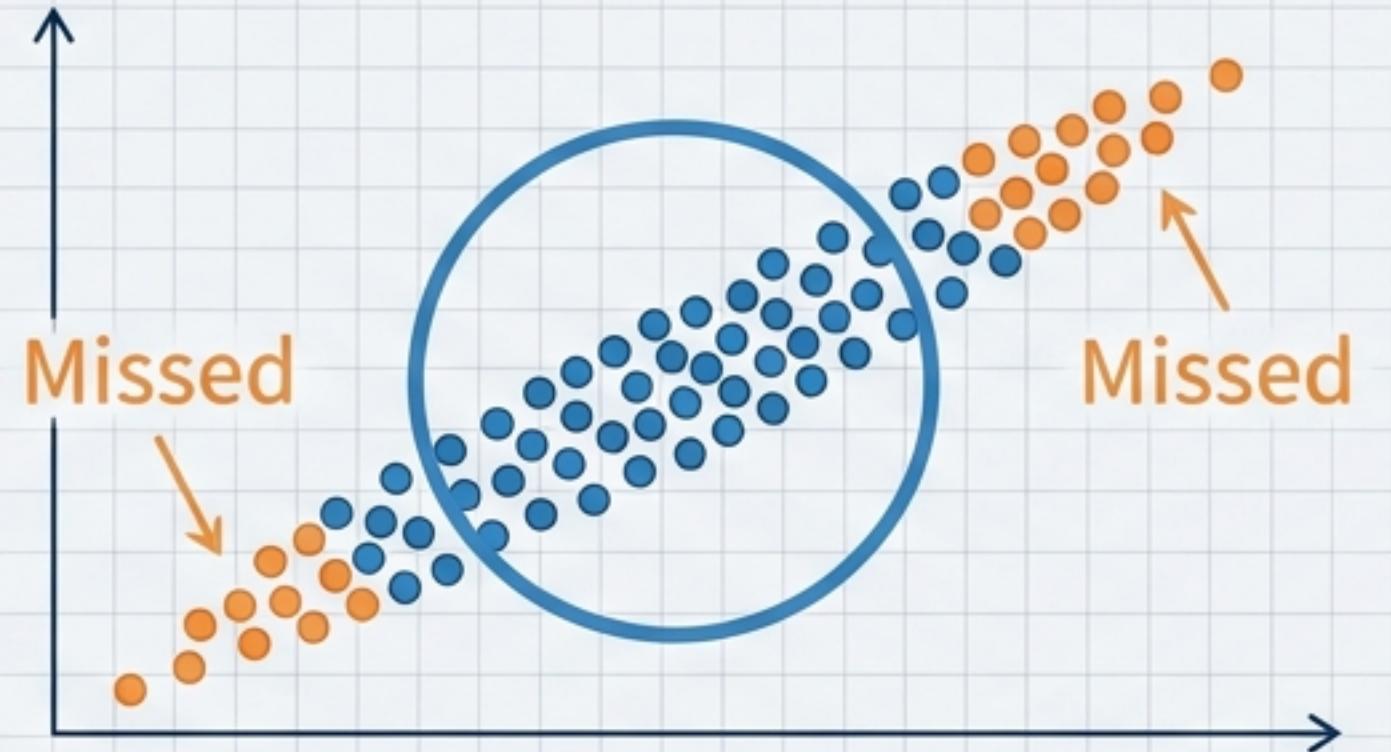
1. 軟分群 (Soft Clustering)：每個數據點都有一定機率屬於每個群集。
2. 混合係數 (π_k)：代表每個群集的權重。
3. 高斯分布 (μ_k, Σ_k)：描述群集的中心位置與形狀範圍。

$$p(x) = \sum_{k=1}^K \pi_k \mathcal{N}(x | \mu_k, \Sigma_k)$$



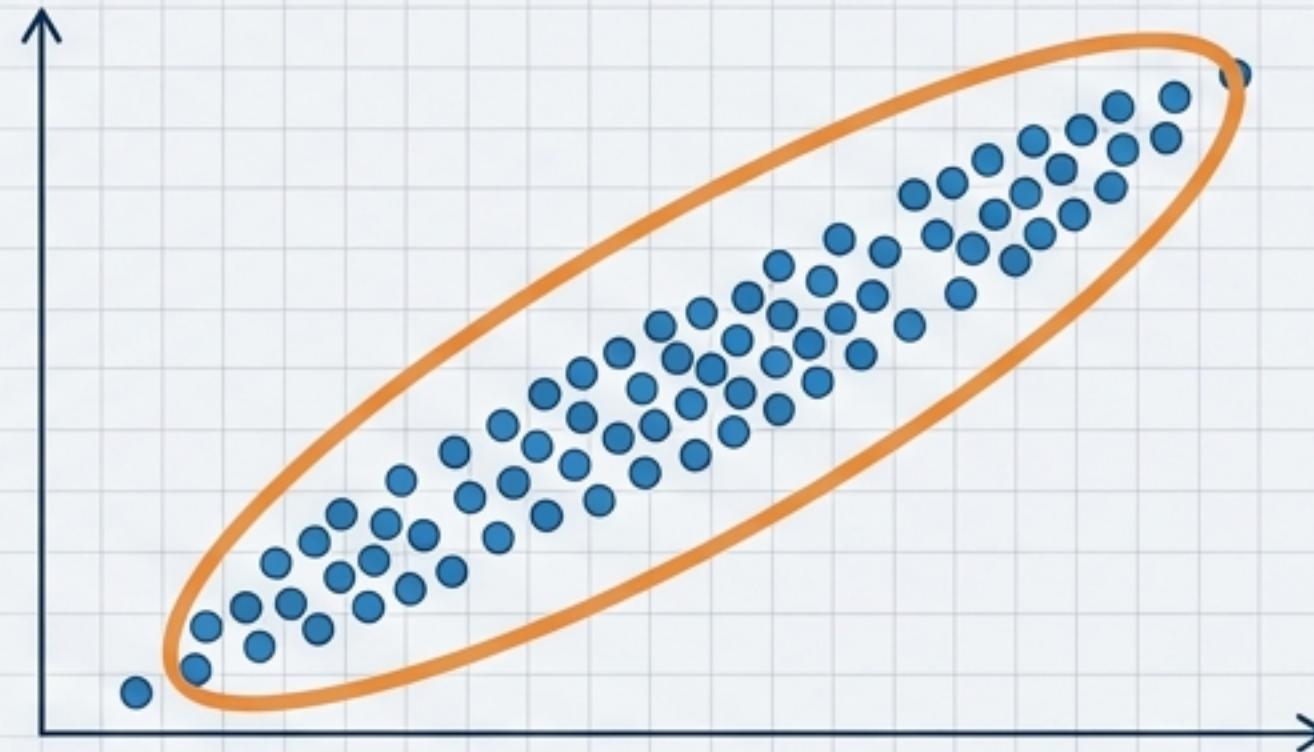
技術對決：GMM vs. K-Means

K-Means (Hard Clustering)



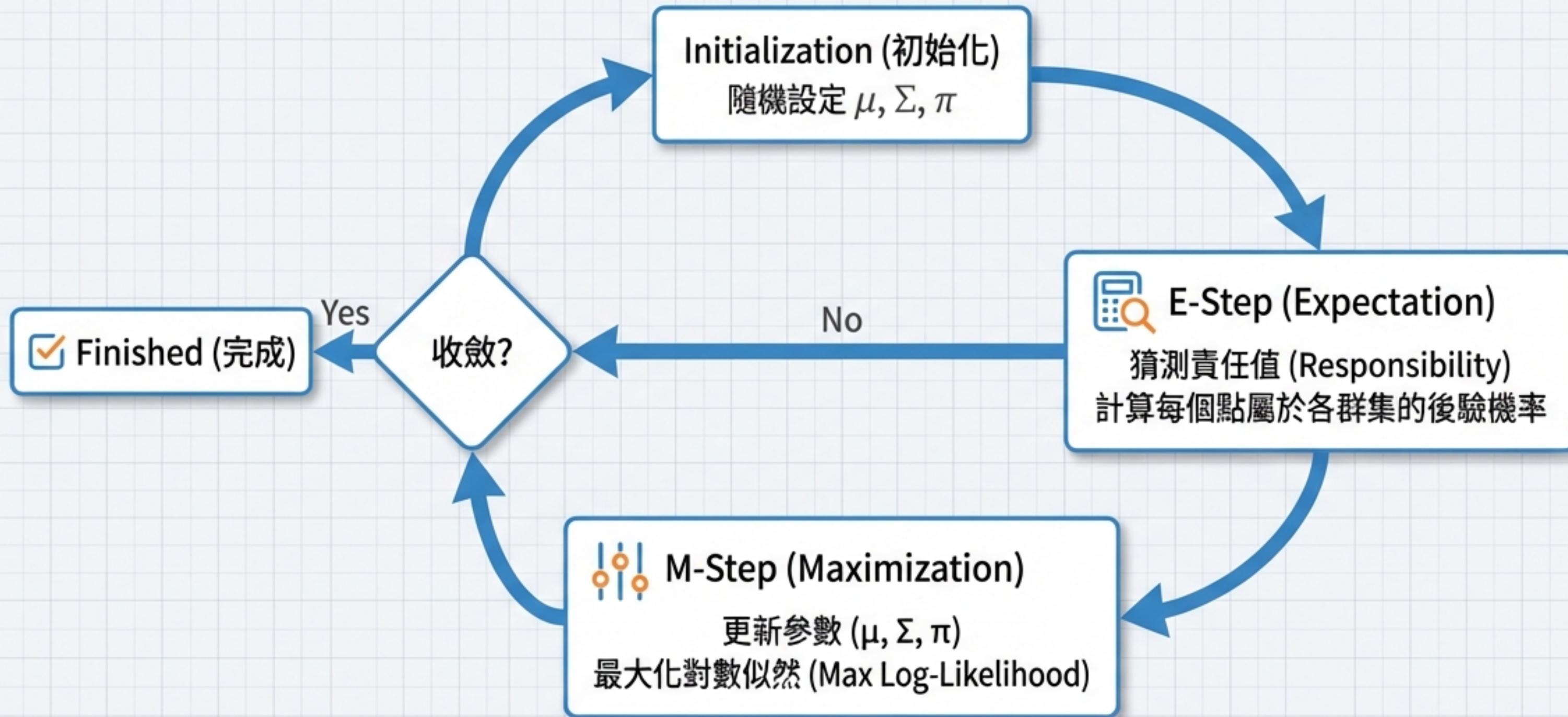
- 硬分群 (是非題)
- 僅限球形 (Spherical)
- 歐幾里得距離

GMM (Soft Clustering)



- 軟分群 (機率題)
- 橢圓形 (可旋轉/拉伸)
- 馬氏距離 (考慮協方差)

學習引擎：期望最大化 (EM) 演算法

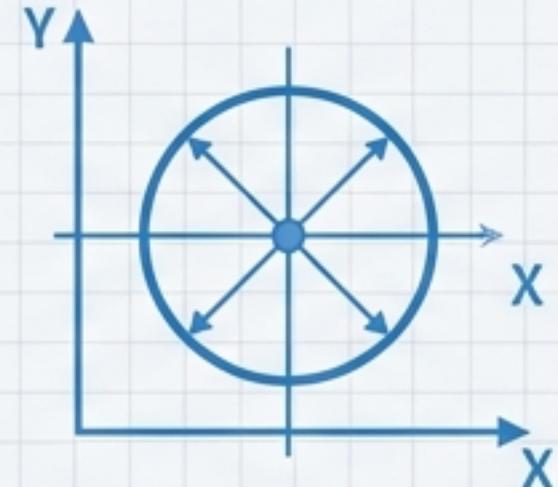


概念類比：類似於化工廠的 PID 控制迴路調整，不斷修正誤差直到系統穩定。

形狀控制：協方差矩陣類型 (Covariance Types)

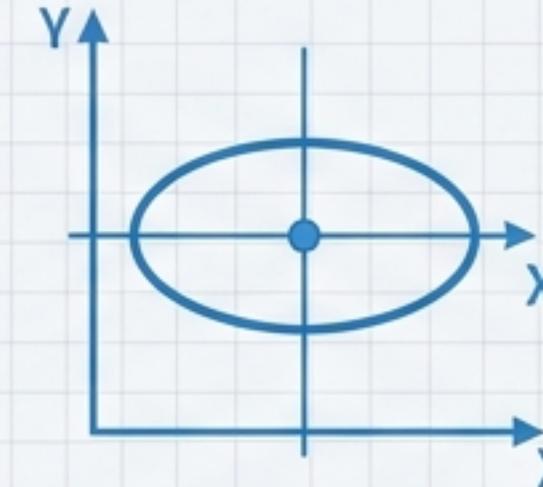
參數 `covariance_type` 決定了群集的幾何自由度

`spherical`



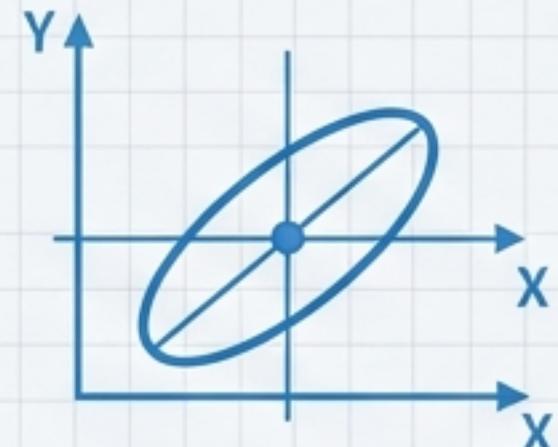
圓形。變數間無相關性，方差相同。

`diag`



軸對齊橢圓。變數間獨立，方差可不同。

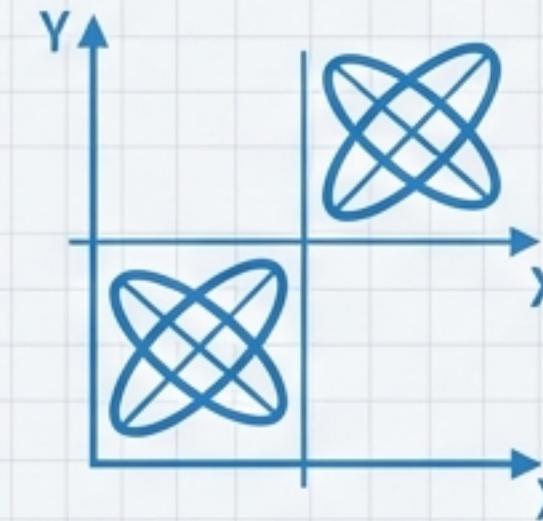
`full`



任意橢圓。變數間有相關性（可旋轉）。

最靈活，擬合化工數據最常用

`tied`



綁定形狀。所有群集共用形狀參數。

模型選擇：尋找最佳的 K 值

資訊準則 (Information Criteria)

不同於 K-Means 的 Elbow Method，GMM 使用更嚴謹的統計指標。

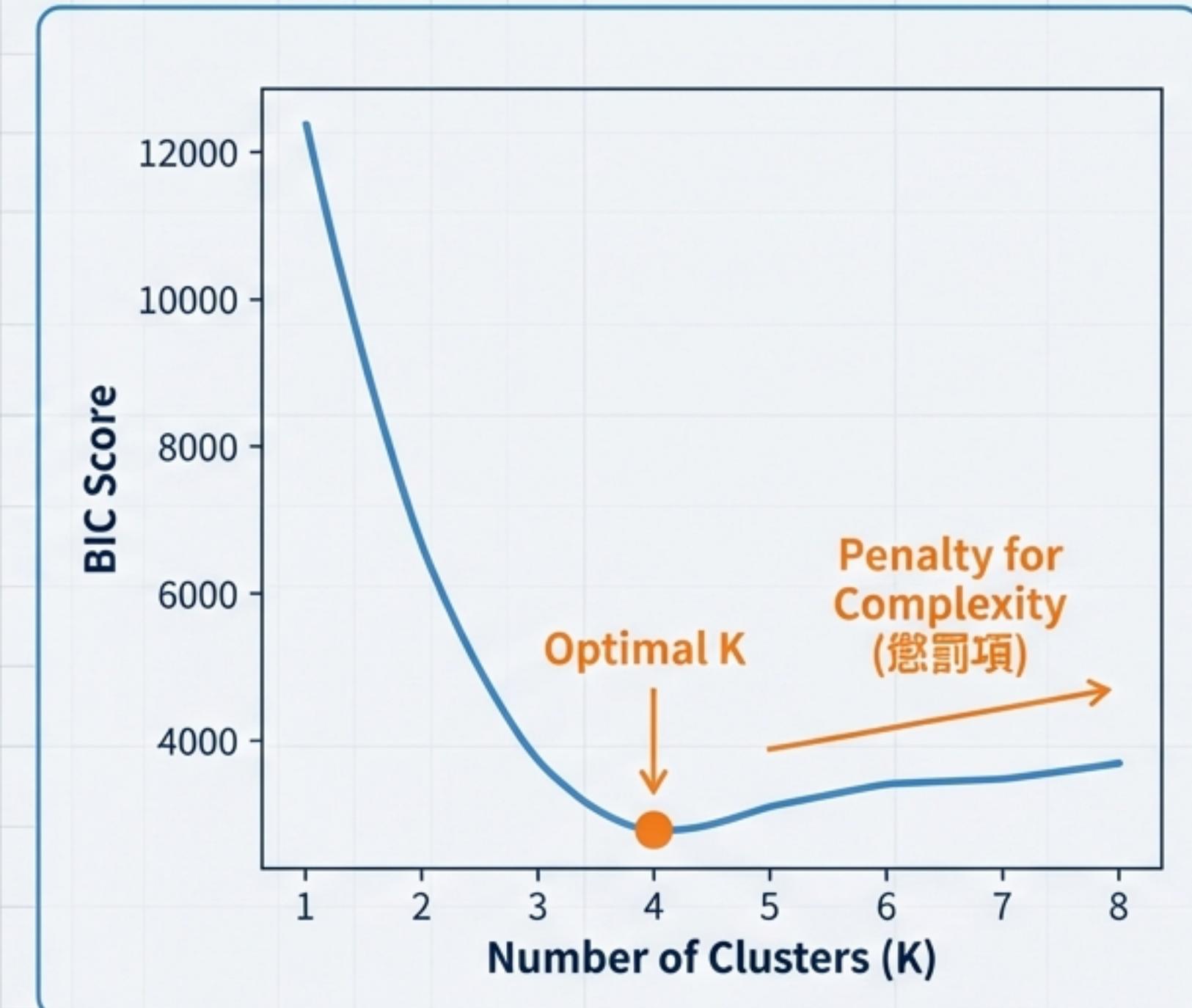
BIC (Bayesian Information Criterion):

$$BIC = -2 \log L + \text{Penalty}$$

- **黃金法則:** 選擇 **BIC 值最小** 的 K
- 特點: 對模型複雜度**懲罰較重**，避免過擬合 (Overfitting)

AIC (Akaike Information Criterion):

- 傾向選擇**較複雜模型**，可作為輔助參考。



實作工具箱：Scikit-learn Implementation

Code Card

```
from sklearn.mixture import GaussianMixture

# 初始化模型
gmm = GaussianMixture(
    n_components=3,           # K: 群集數量 (需透過 BIC 選擇)
    covariance_type='full',   # 允許橢圓形狀 (適合化工數據)
    n_init=10,                # 隨機初始化次數 (避免局部最優)
    random_state=42
)

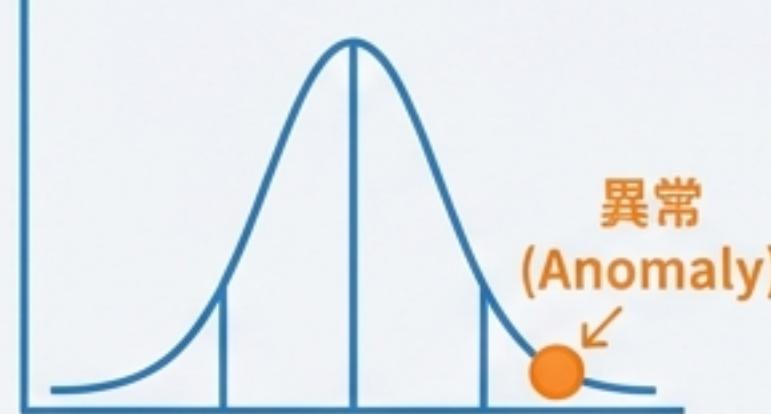
# 訓練與預測
gmm.fit(X_train)
labels = gmm.predict(X_test)      # 硬分群結果
probs = gmm.predict_proba(X_test) # 軟分群機率 (關鍵功能)
```

最關鍵參數
(Critical Parameter)

確保收斂穩定性
(Stability)

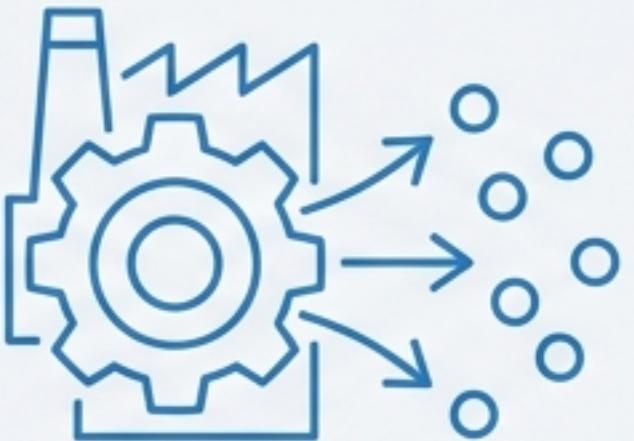
獲取不確定性資訊
(Uncertainty)

GMM 的獨特優勢：生成與檢測



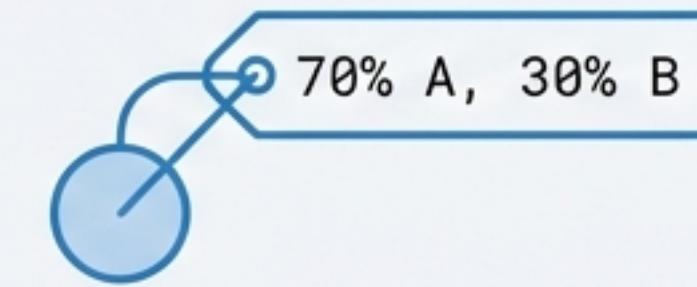
異常檢測 (Anomaly Detection)

- 計算對數似然
(Log-Likelihood)
- 低機率密度區域 = 潛在故障或異常



數據生成 (Generative Model)

- `gmm.sample(n)`
- 生成合成數據以增強資料集
- 模擬各種製程情境

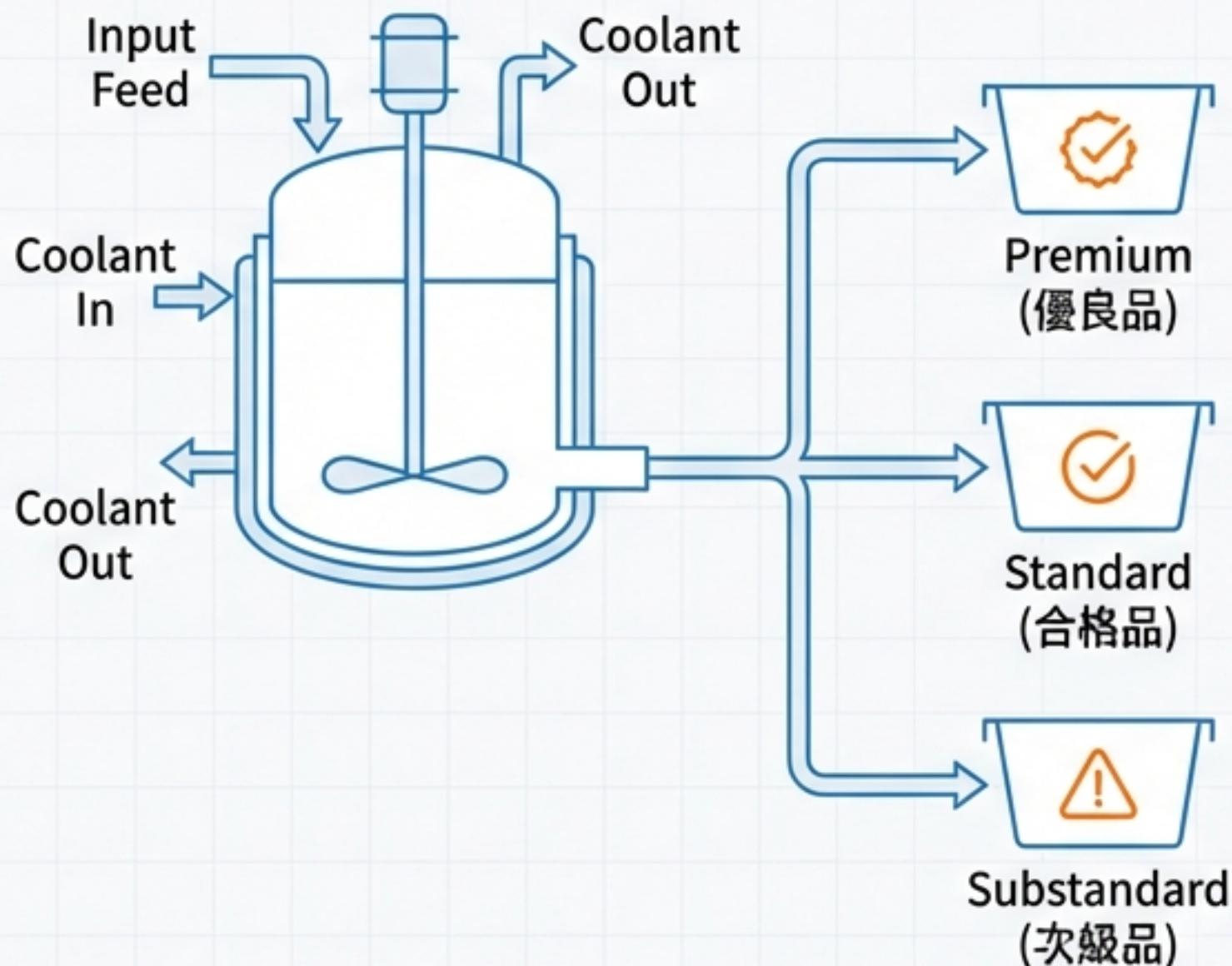


不確定性量化 (Uncertainty)

- 軟分群核心價值
- 明確給出每個點的歸屬機率
- 支援風險評估矩陣

實戰案例：反應器多產品品質分布建模

CSTR Reactor & Product Distribution



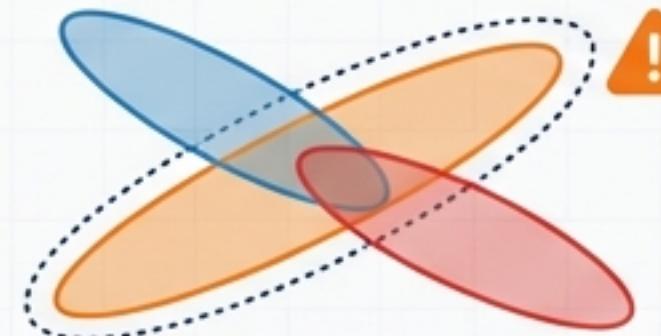
Data & Modeling Challenge

數據特徵 (Data Features)

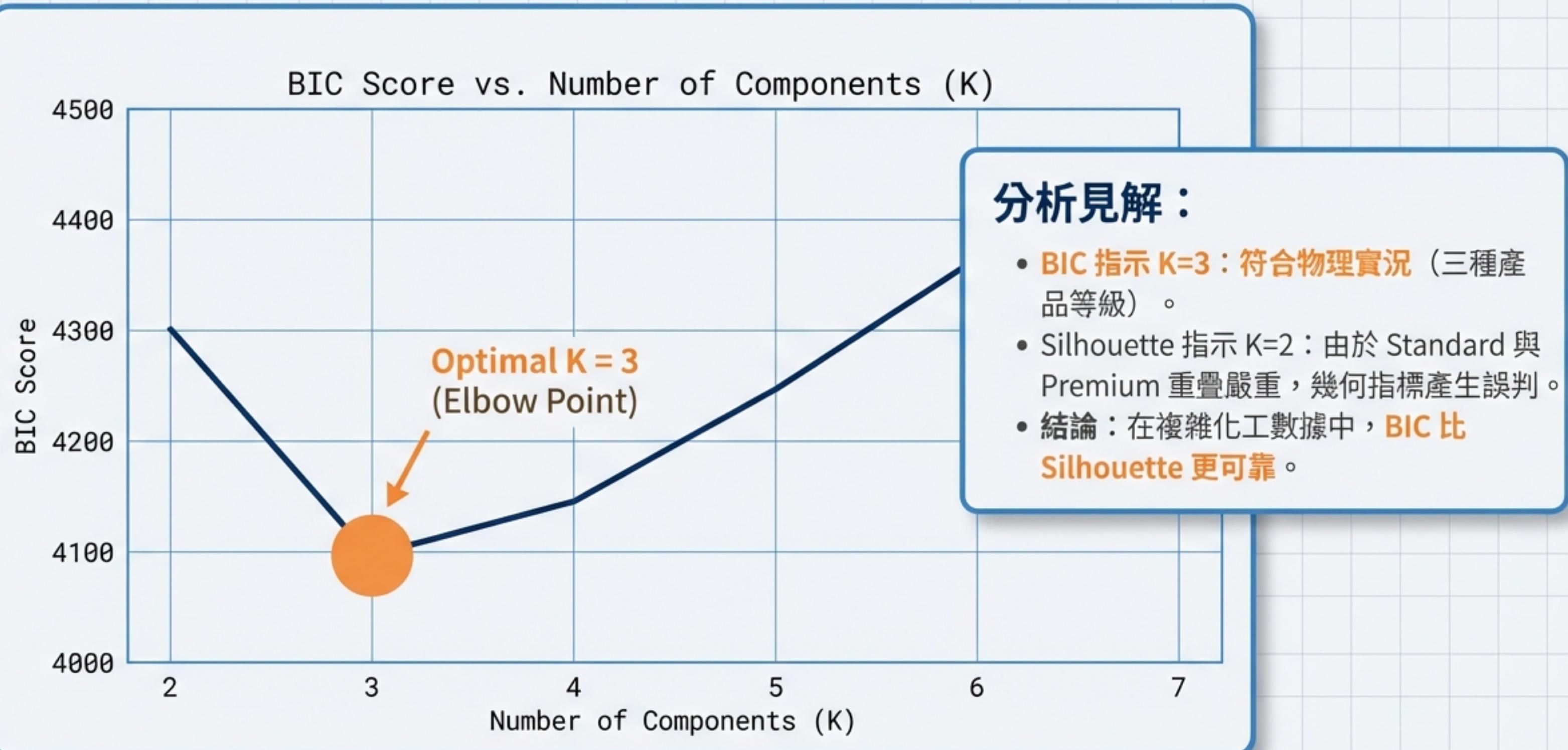
- Yield (產率)
- Purity (純度)
- Selectivity (選擇性)

挑戰 (The Challenge)

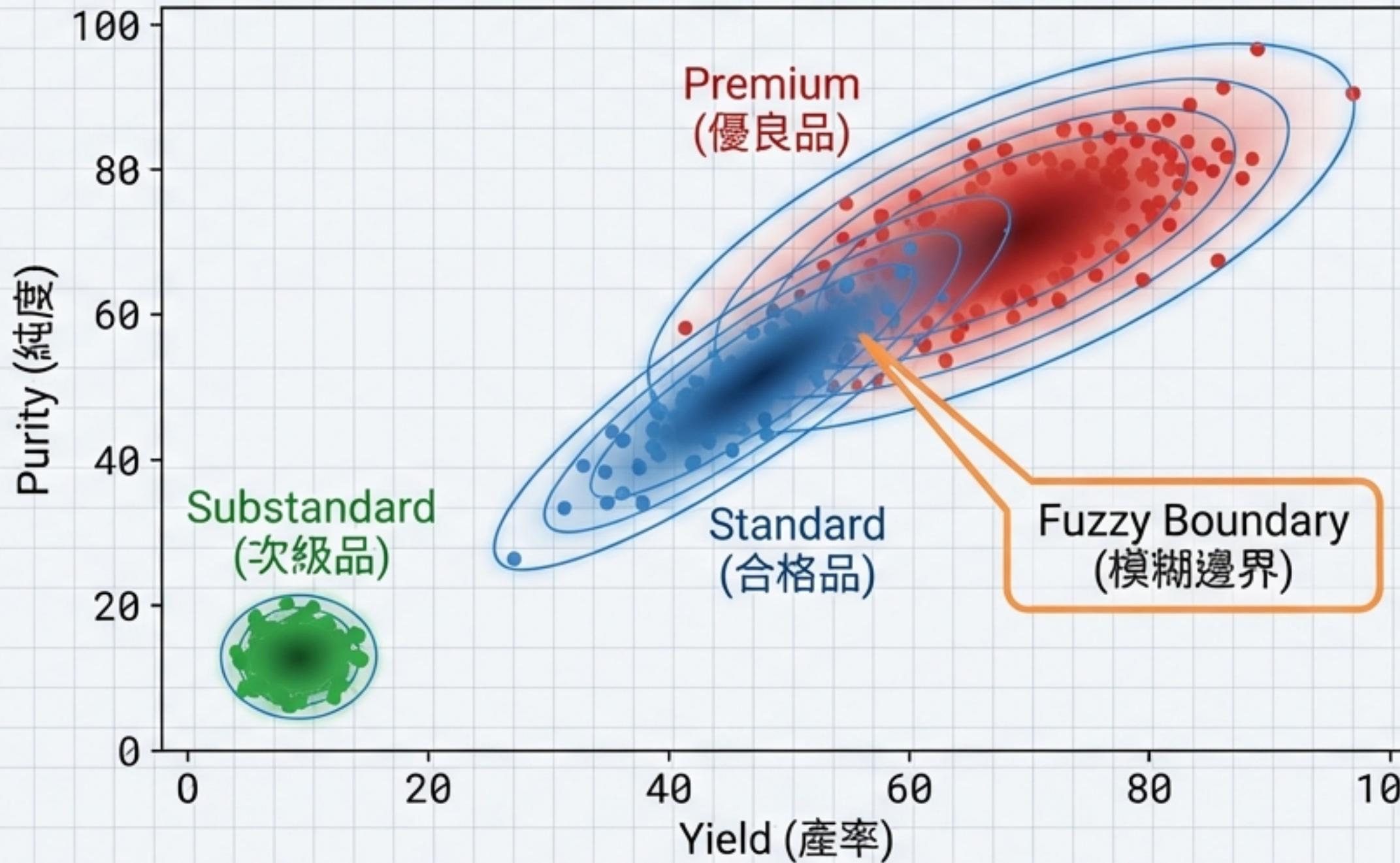
數據存在顯著的重疊區域與相關性 (橢圓形分布)。傳統硬分群難以區分處於「過渡帶」的產品品質。



數據驅動決策：決定群集數量



透視模糊邊界：軟分群結果視覺化



價值轉化：風險評估與決策支援系統

Case Example: Borderline Product #4

Model Output
Probabilities:

- Standard: 52%
- Premium: 41%
- Substandard: 7%

Risk Matrix Assessment

Low Risk (Prob > 70%)	Auto-Approve (自動放行)
Medium Risk (30-70%)	Manual Check (人工複檢)
High Risk (Prob < 30%)	Anomaly Alert (異常警報)

Decision (決策)



**Flag as
Medium
Risk**

避免將潛在 Premium 產
品誤判為 Standard，挽
回經濟價值。

GMM 在化工領域的多元應用



反應器操作模式識別

識別開車、停車、穩態操作之間的過渡狀態。



溶劑篩選 (Solvent Screening)

評估候選溶劑屬於各性能群組的機率，降低研發風險。



批次製程 (Batch) 監控

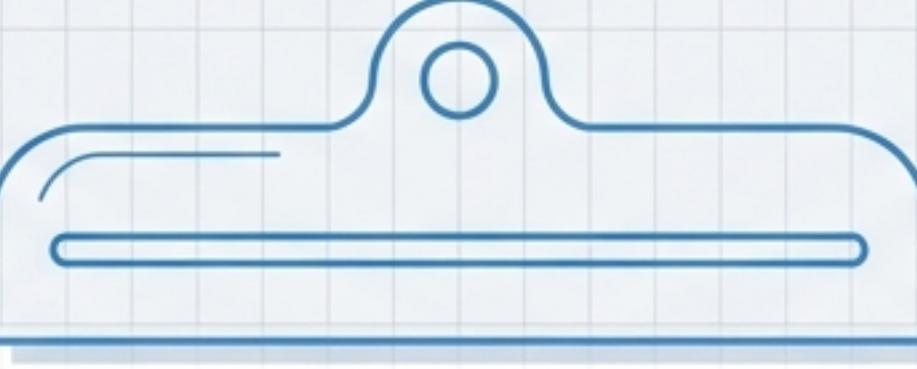
比較新批次軌跡與歷史黃金批次 (Golden Batch) 的相似度。



異常檢測 (Anomaly Detection)

利用對數似然值監控感測器數據，提前預警設備故障。

最佳實踐與總結 (Best Practices)

- 
- 數據預處理：務必進行標準化 (Standardization)，因 GMM 對尺度敏感。
 - 模型選擇：使用 BIC 尋找最佳 K 值，而非僅依賴 Elbow Method。
 - 參數設定：複雜數據使用 `covariance_type='full'`，並設定 `n_init > 10`。
 - 應用場景：當需要「不確定性量化」與「風險評估」時，GMM 優於 K-Means。

Bottom Line: GMM 將模糊的『可能』轉化為精確的『機率』，為工程決策提供數據支撐。