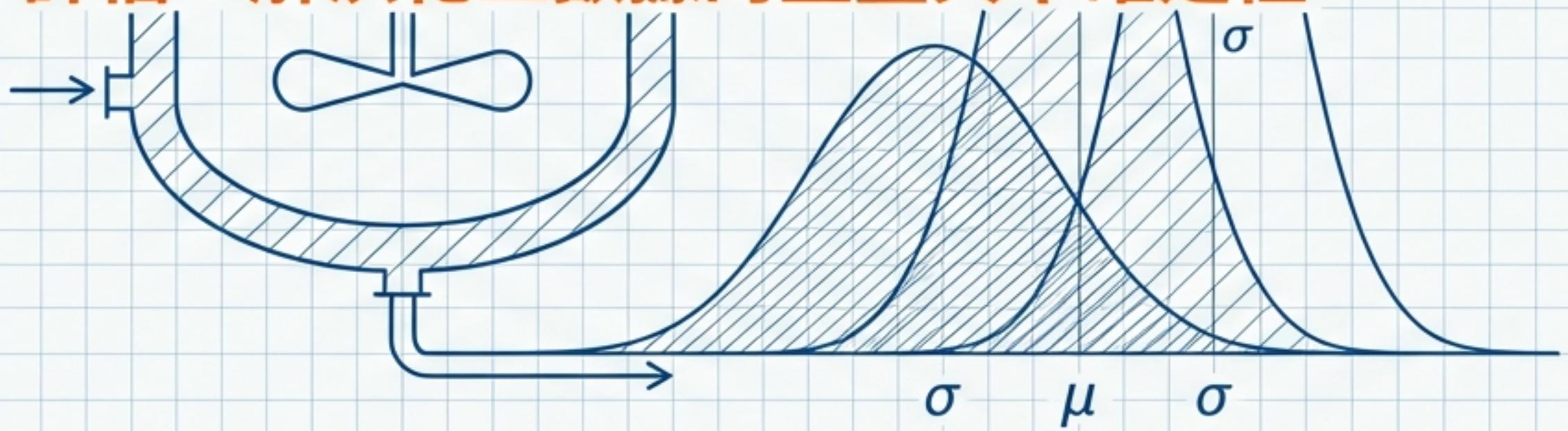


Unit 05 高斯混合模型 (Gaussian Mixture Models)

從硬分群到機率評估：解決化工數據的重疊與不確定性



課程目標：掌握機率分群的核心技術

Roboto Mono

 理解原理：掌握 GMM 的機率生成機制與高斯分布混合概念。

 掌握算法：理解期望最大化（EM）演算法的迭代優化過程。

 實戰應用：學會使用 scikit-learn 進行反應器產品品質分級與風險評估。

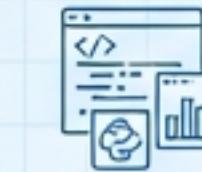
 模型決策：利用 BIC 與 AIC 準則選擇最佳群集數量 (K)。

Roboto

Course Process Map

Phase 1: 基礎 (Foundation)

Part 0-1: Python, Numpy, Pandas, Matplotlib



Phase 2: 非監督式學習 (Unsupervised)

分群，降維，異常檢測



Units 05-09

Phase 3: 監督式學習 (Supervised)

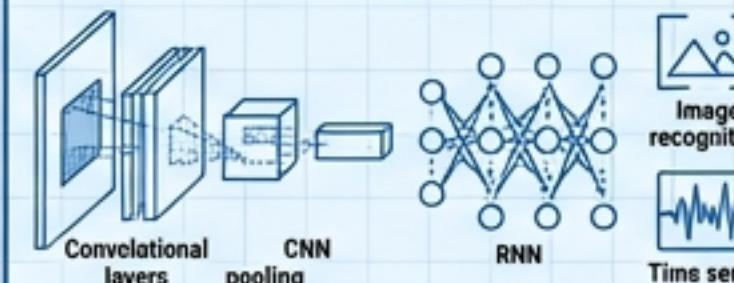
回歸，分類，模型評估



Units 10-14

Phase 4: 深度學習 (Deep Learning)

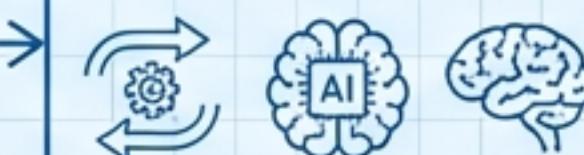
CNN, RNN, GRU, 應用案例



Units 15-18

Phase 5: 進階 (Advanced)

RL, GenAI, LLMs



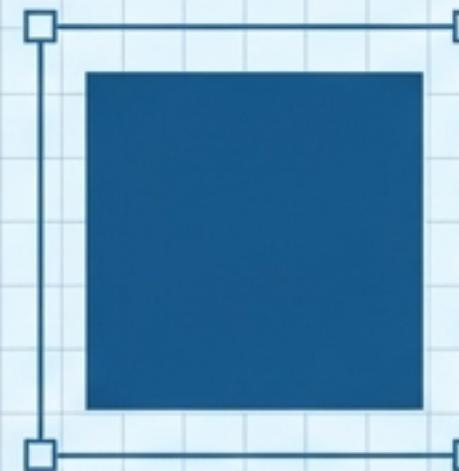
建議初學者從基礎 Python 開始；有經驗者可直接進入 Part 2。

You Are Here

分群思維的轉變：硬分群 vs. 軟分群

Roboto Mono

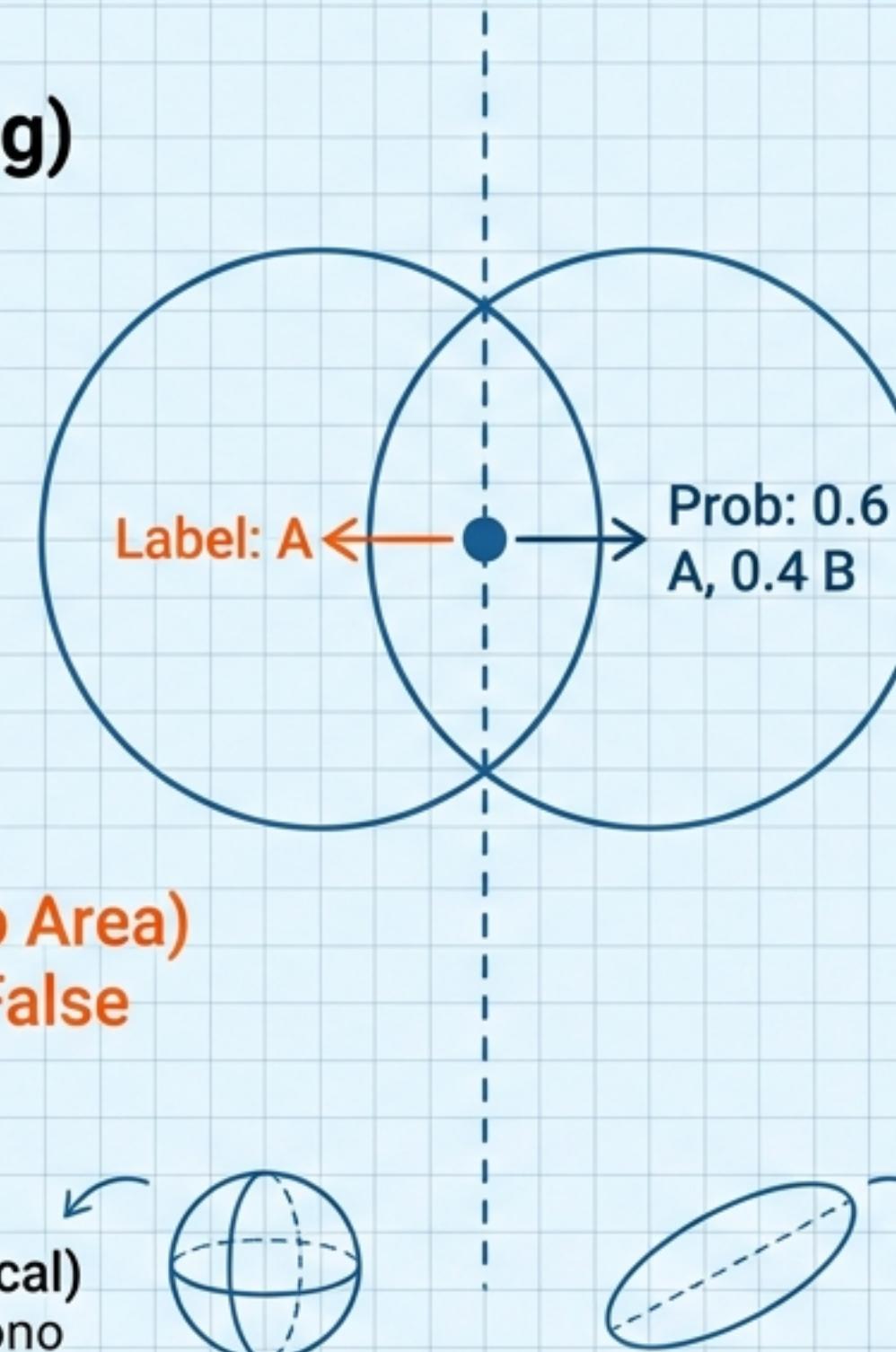
K-Means (Hard Clustering)



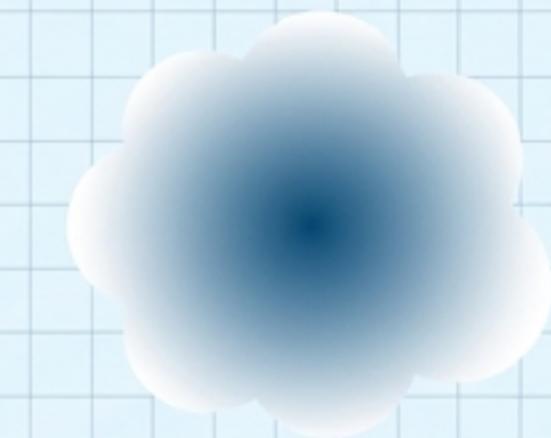
每個數據點只能屬於一個群集
(0 或 1)。

缺陷：在群集重疊區域 (Overlap Area)
會強制分類，導致錯誤的自信 (False
Confidence)。

幾何假設：球形 (Spherical)
Roboto Mono



GMM (Soft Clustering)



每個數據點擁有屬於各群集的
機率 (e.g., 70% A, 30% B)。

優勢：量化不確定性 (Quantify
Uncertainty)，保留邊界資訊。

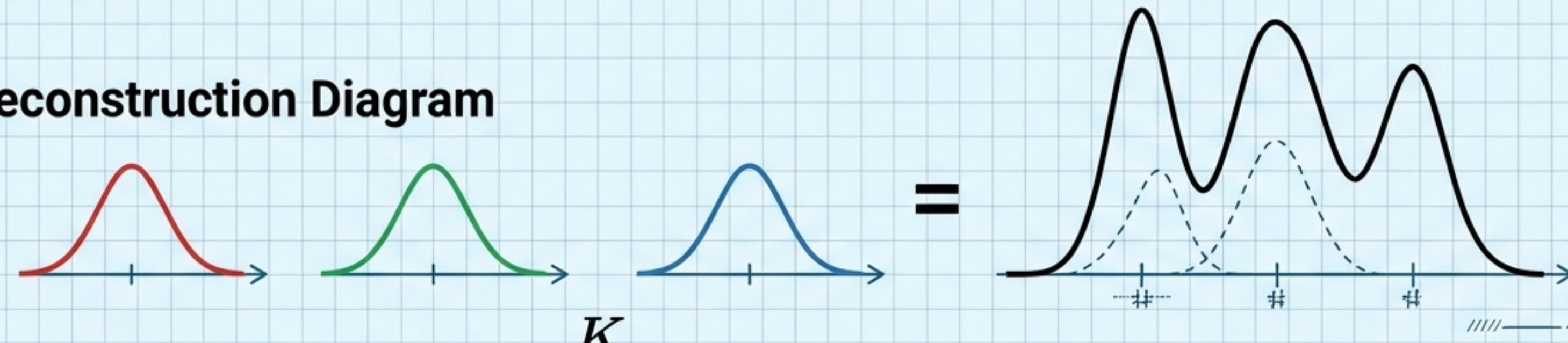


幾何假設：橢圓形 (Elliptical)
Roboto Mono

解密 GMM：數據是由高斯分布混合生成的

Roboto Mono

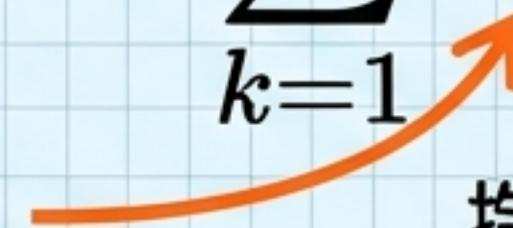
Deconstruction Diagram



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



混合係數 (Weight)
- 操作模式的相對比重



均值 (Mean)
- 操作模式的中心點 (Setpoint)

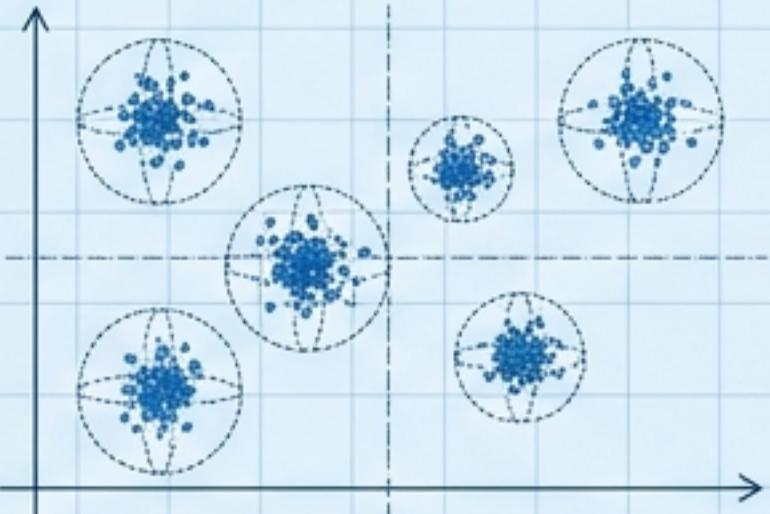


協方差 (Covariance)
- 變數間的相關性與形狀

工程類比：想像反應器有多個操作模式 (Operating Modes)，最終的產出數據是這些模式的混合結果。

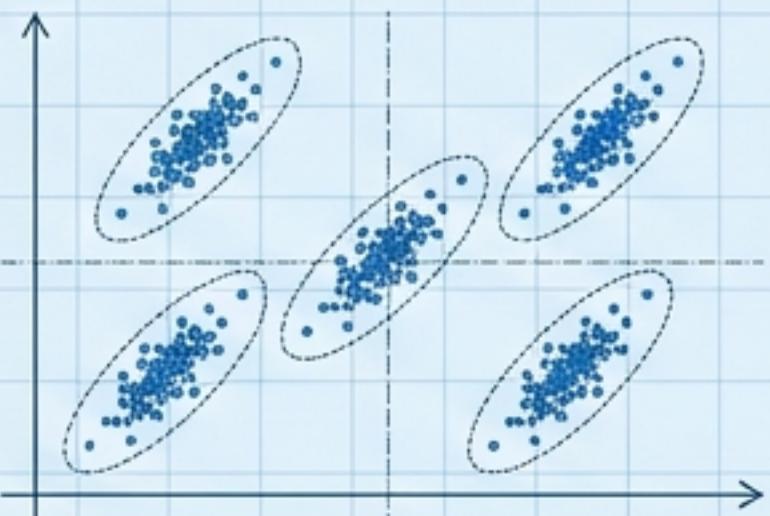
適應數據的形狀：協方差矩陣 (Covariance Matrix)

Spherical (球形)



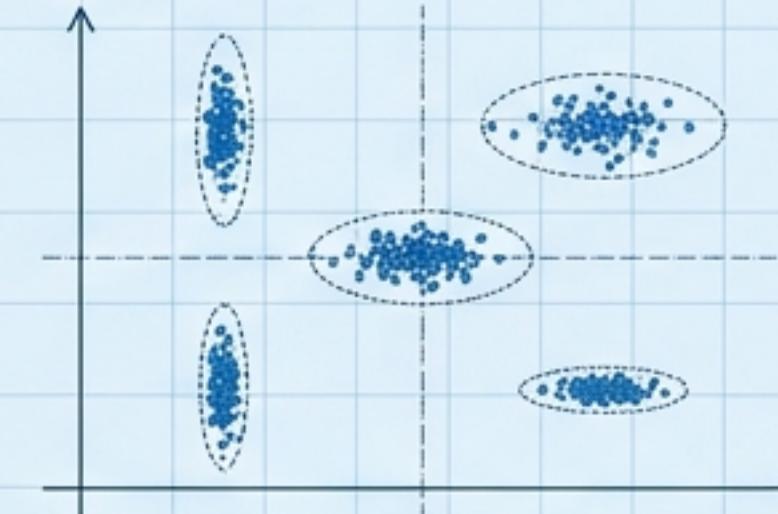
類似 K-Means。假設各方向變異相同，特徵獨立。

Tied (綁定)



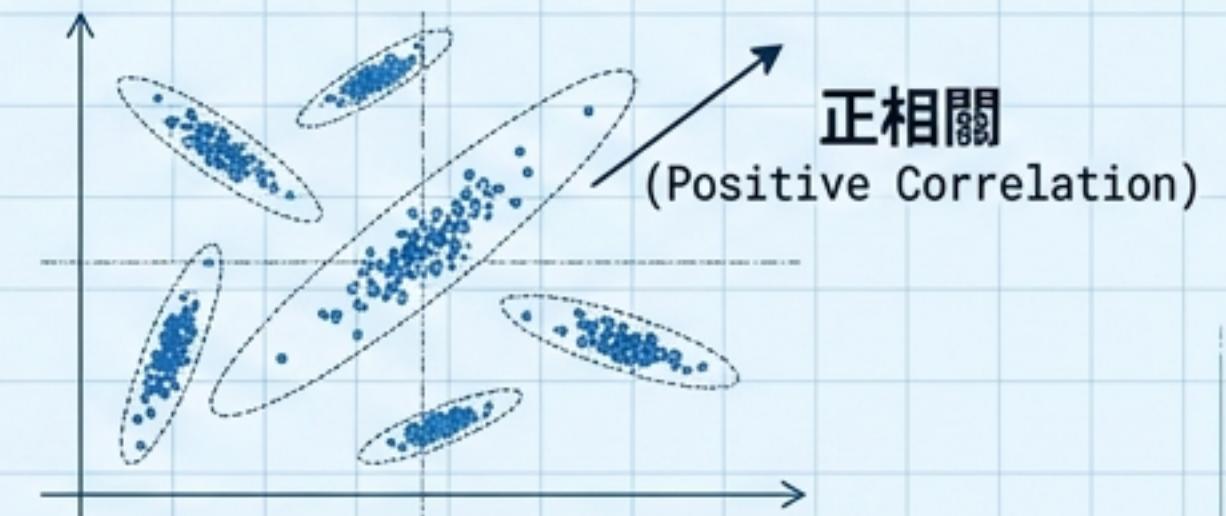
所有群集共享形狀參數。

Diagonal (對角)



特徵間無相關性 (Independent features)。

Full (完全) - 化工首選

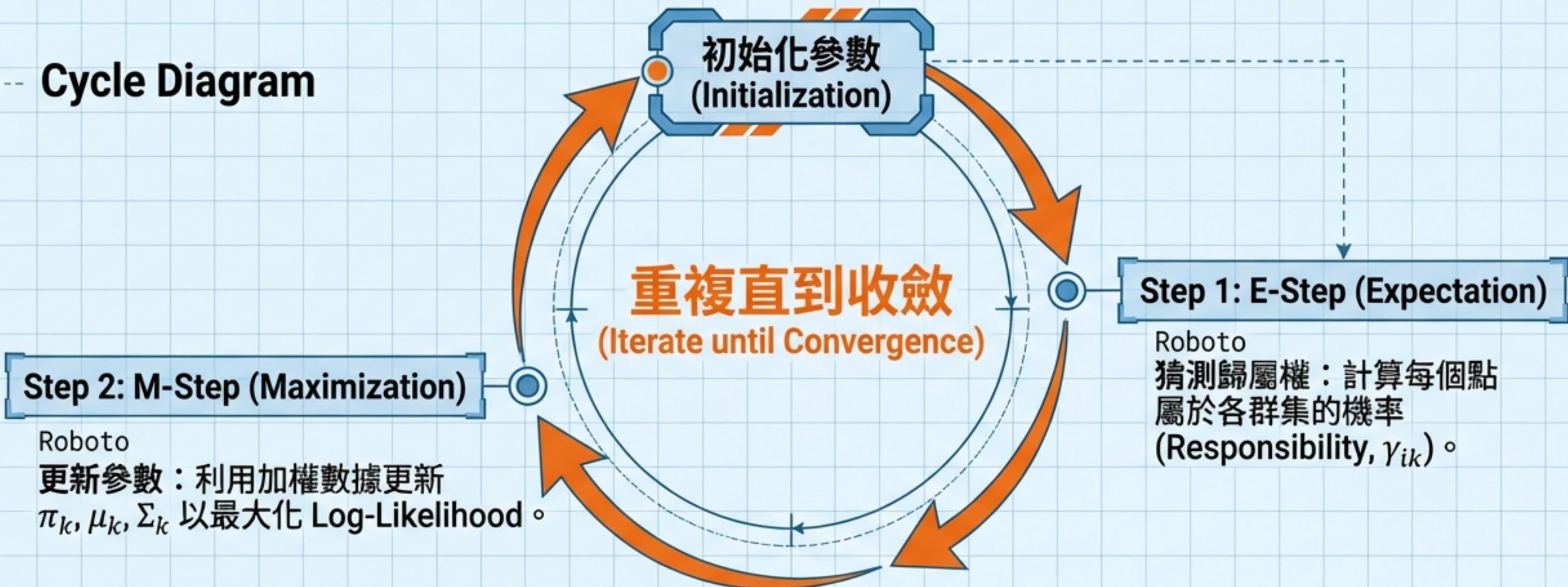


能捕捉變數間的相關性 (Correlation)。
例如：溫度升高導致壓力升高 (正相關)。

參數估計：期望最大化 (EM) 演算法

Roboto Mono 迭代優化：在未知中尋找最佳解

Cycle Diagram



注意：EM 保證收斂到局部最優 (Local Optimum)，實務上需設定 $n_init > 1$ 進行多次嘗試。

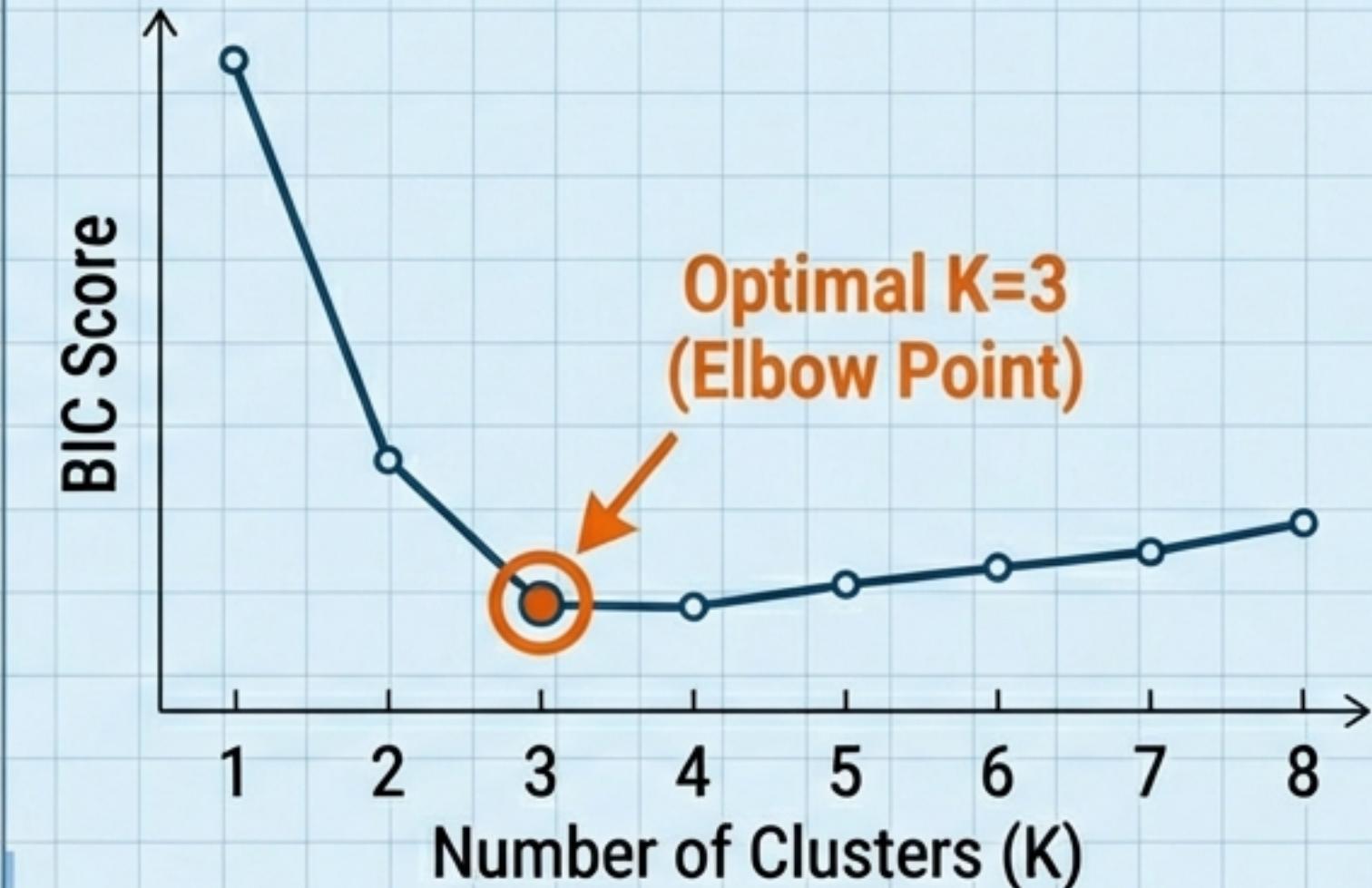
決策關鍵：如何選擇最佳群集數量 K？

Roboto Mono

衝突點 (The Conflict)

- Silhouette Score：偏好分離明確的群集，但在重疊數據上往往失效。
- BIC (Bayesian Information Criterion)：平衡擬合度與模型複雜度，是 GMM 的黃金標準。

BIC Score Evaluation



法則：選擇 BIC 最小值。並結合化工領域知識（例如：已知產品分為 優/良/劣 3 級）。



實戰案例：反應器多產品品質分布建模

Roboto Mono

設備：連續式攪拌槽反應器 (CSTR)

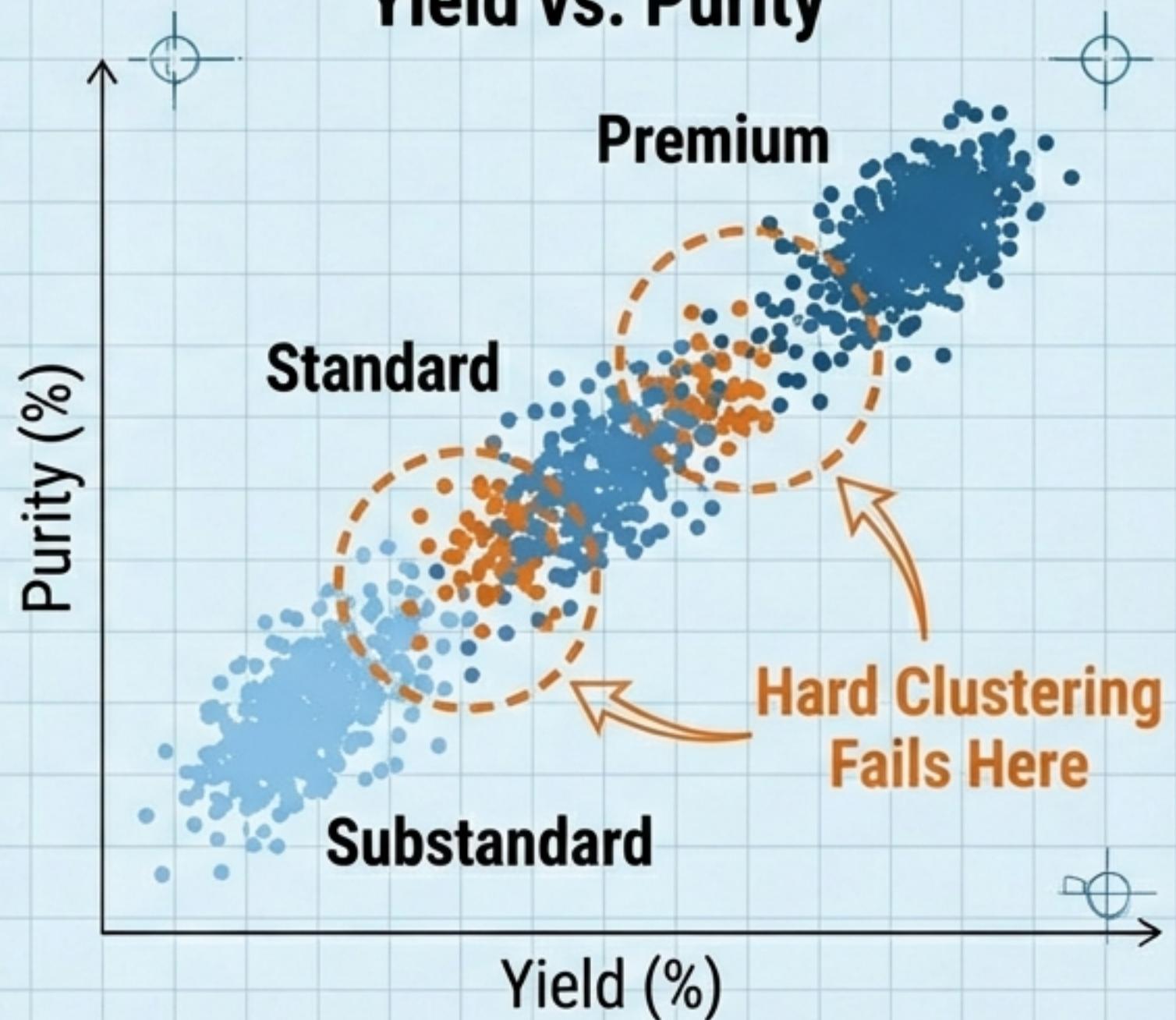
產品分級：3 Grades

1. Premium (優良)
2. Standard (合格)
3. Substandard (次級)

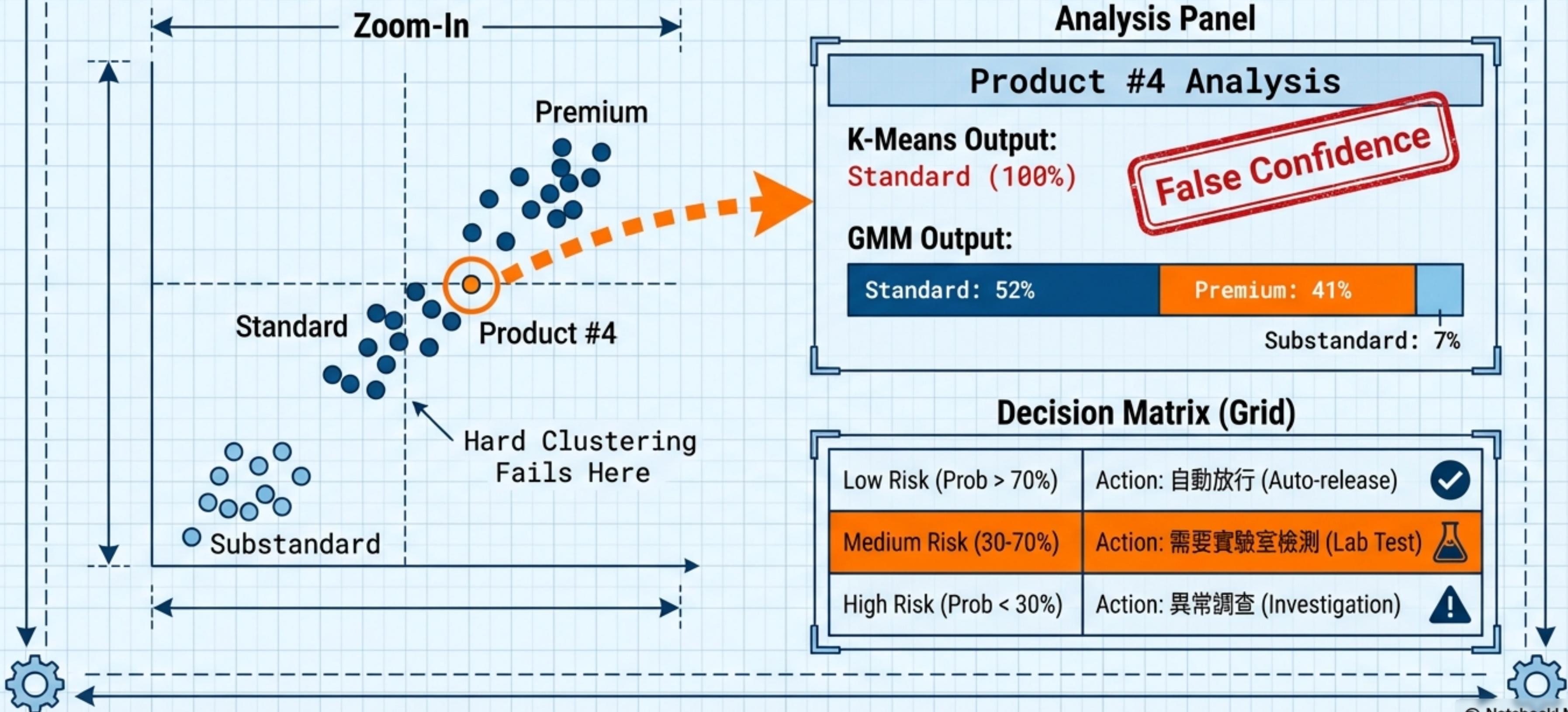
挑戰：

“Standard” 產品位於中間過渡帶，
與其他兩者重疊。

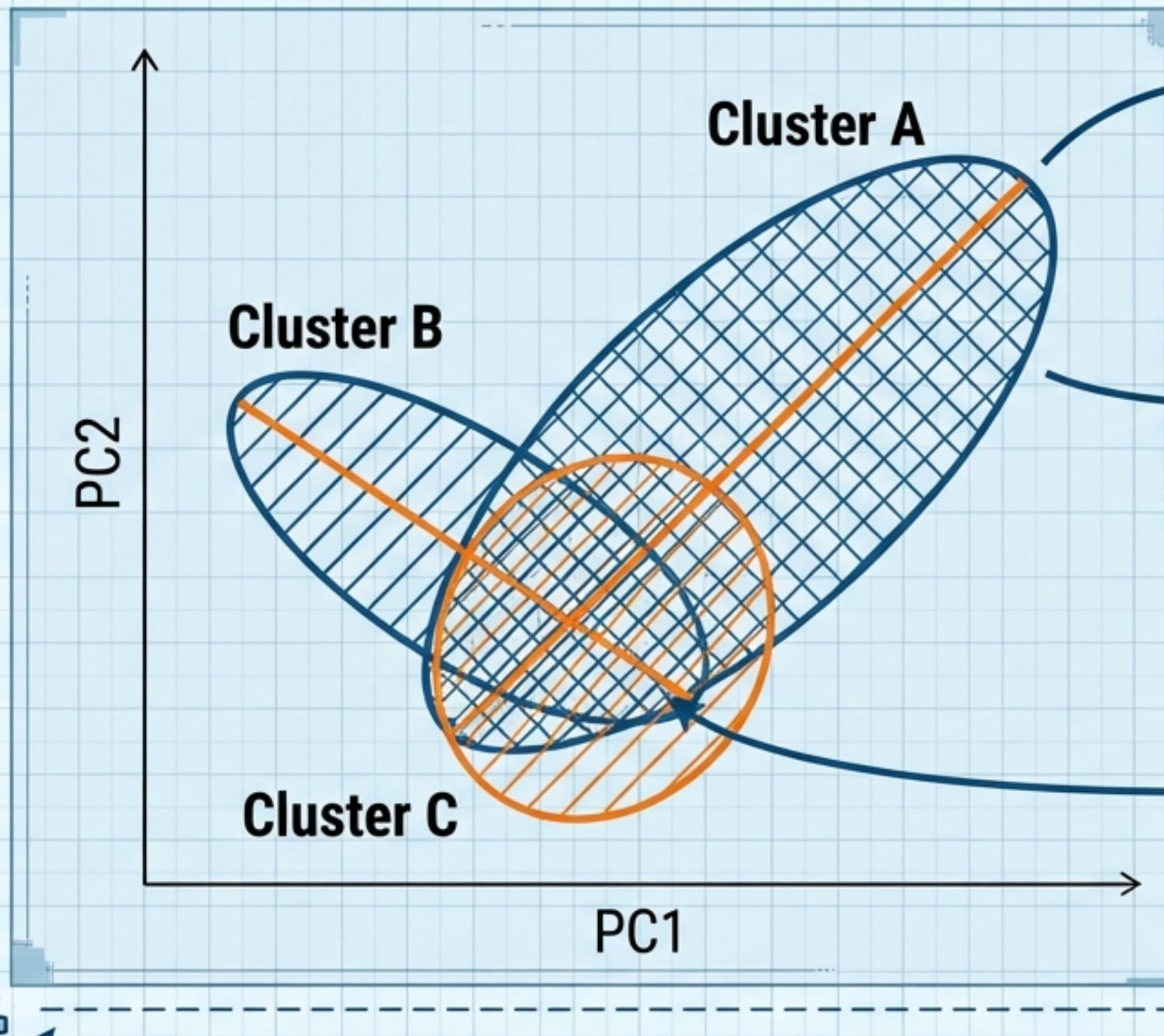
Yield vs. Purity



GMM 的價值：量化不確定性與風險評估



物理意義視覺化：協方差橢圓 (Covariance Ellipses)



方向 (Orientation) = 變數相關性
例如：產率與純度的 Trade-off 關係。

大小 (Size) = 製程穩定性
大橢圓表示變異 (Variance) 較大，
製程控制較鬆散。

重疊 (Overlap) = 過渡狀態
代表反應器處於操作模式切換的過渡期。

AI 在化工全生命週期的價值 (GMM 應用版)



識別反應器是否處於
穩態或過渡態。

利用 Log-Likelihood。
若某點在所有群集的
機率都極低 -> 漏漏或
故障。

根據物理性質分群，
評估新溶劑屬於各性
能群組的機率。

利用軟感測器預測產
品品質機率，減少離
線分析。

實作流程：使用 Python scikit-learn

```
from sklearn.mixture import GaussianMixture

# 1. 初始化模型
gmm = GaussianMixture(
    n_components=3,          # K=3 (由 BIC 決定)
    covariance_type='full',   # 關鍵：捕捉化工變數相關性
    n_init=10,                # 多次初始化避免局部最優
    random_state=42
)

# 2. 訓練與預測
gmm.fit(X_train)
probs = gmm.predict_proba(X_test) # 取得機率矩陣
```

關鍵參數：必須選擇
'full' 以適應化工數
據的橢圓形分布特性。

選擇指南：GMM vs. K-Means

特性 (Feature)

K-Means

GMM (本單元重點)

幾何形狀

球形 (Spherical)



輸出結果

硬標籤 (Label)



計算速度

快 (Fast)



重疊處理

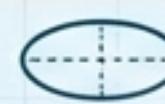
差 (強制切割)



適用場景

快速初探、分離明確數據

橢圓形 (Elliptical)



機率分布 (Probability)



較慢 (Slower)



優 (軟分群)

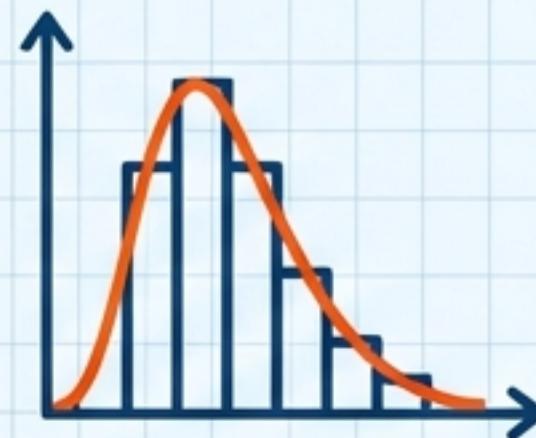


風險評估、相關性數據

Verdict：若需要精確的邊界處理與風險量化，GMM 是化工應用的首選。

現實世界的挑戰與最佳實踐

非高斯分布數據



Problem (問題)：數據呈現偏態 (Skewed)。

Solution (解決方案)：使用 Log 轉換或 Box-Cox 轉換使其接近常態分布。



初始化敏感

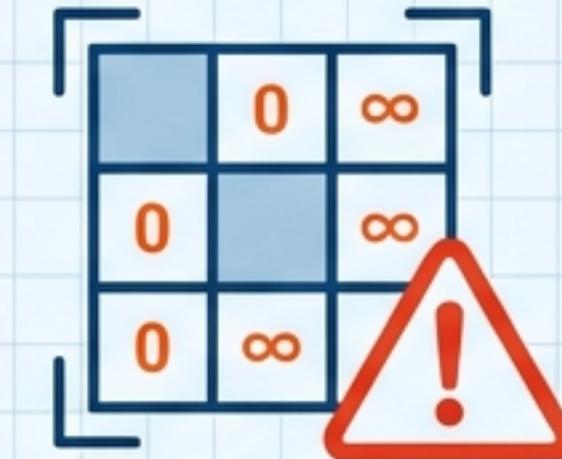


Problem (問題)：EM 演算法可能陷入局部解。

Solution (解決方案)：使用 `init_params='kmeans'` 並設定 `n_init >= 10`。

「 使用 `init_params='kmeans'` 並設定 `n_init >= 10`。」

奇異矩陣 (Singular Matrix)



Problem (問題)：維度過高或樣本過少導致矩陣不可逆。

Solution (解決方案)：使用 PCA 降維或增加正則化項 (`reg_covar`)。



結語：您是未來的定義者

軟分群不是不精確，而是更誠實地反映物理世界的不確定性。

下一步 (Next Step)

前往 Jupyter Notebook：
`Unit05_Gaussian_Mixture_Models.ipynb`
動手實作：反應器品質數據分析與風險評估系統。

