

# 基于高斯混合模型的非参数控制图

唐金会

成都理工大学数学科学学院, 四川 成都

收稿日期: 2026年3月2日; 录用日期: 2026年3月26日; 发布日期: 2026年4月7日

## 摘要

在工业生产过程中, 由于设备参数、原料特性以及工艺条件的异质性, 过程数据往往呈现出复杂的多模态分布特征。这类数据结构超出传统统计过程控制方法所依赖的单模态或同分布假设的适用范围。多模态数据复杂的多峰特性, 使得基于参数化分布假设的传统建模方法难以准确刻画数据的真实分布, 进而影响过程监控。针对这种情况, 本文提出了一种数据驱动的监控方法。首先, 引入高斯混合模型对多模态过程数据进行概率密度建模, 克服了传统方法在分布形式上的先验假设局限; 其次, 基于该模型构建负对数似然统计量, 并结合指数加权移动平均策略, 实现对过程异常的检测。数值仿真与实例验证了本文所提控制图的可行性和实用价值, 为多模态过程的质量监控提供了一种有效的新思路。

## 关键词

高斯混合模型, 多模态数据, 非参数控制图, 统计过程控制

# Nonparametric Control Chart Based on Gaussian Mixture Model

Jinhui Tang

School of Mathematical Sciences, Chengdu University of Technology, Chengdu Sichuan

Received: March 2, 2026; accepted: March 26, 2026; published: April 7, 2026

## Abstract

In industrial production processes, process data often exhibit complex multimodal distribution characteristics due to heterogeneity in equipment parameters, raw material properties, and process conditions. Such data structures exceed the applicability of traditional statistical process control methods, which typically rely on unimodal or identically distributed assumptions. The intrinsic multi-peak nature of multimodal data makes it difficult for conventional parametric modeling approaches to accurately capture the true data distribution, thereby compromising process monitoring performance.

To address this issue, this paper proposes a data-driven monitoring method. First, a Gaussian mixture model is introduced to model the probability density of multimodal process data, overcoming the limitations of prior distributional assumptions inherent in traditional methods. Based on this model, a negative log-likelihood statistic is constructed and integrated with an exponentially weighted moving average strategy to detect process anomalies. Numerical simulations and real-world case studies demonstrate the feasibility and practical value of the proposed control chart, offering an effective new approach for quality monitoring in multimodal processes.

## Keywords

Gaussian Mixture Model, Multimodal Data, Nonparametric Control Chart, Statistical Process Control

Copyright © 2026 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

## 1. 引言

统计过程控制(SPC)作为一种过程稳定性监控技术,已在工业制造领域得到广泛应用。传统 SPC 控制图(例如 Shewhart  $\bar{X}$  图)的理论基础建立在一个关键假设之上,受控过程数据服从单模态概率分布,即其概率密度函数(Probability Density Function, PDF)呈现单一峰值特征[1]。然而,实际工业过程往往表现出显著的时变特性,包含动态和非线性等非平稳行为特征[2]。当生产过程涉及多种操作模式切换时,采集到的过程数据可能呈现多模态分布特性,即其 PDF 中存在多个峰值。在本文中,这种由多峰分布所生成的过程数据将被称为多模态数据。

多模态过程数据在复杂工业系统中都有出现,由于这类数据不满足单模态和同分布的基本假设,因此难以用某一具体的参数分布来准确建模。多模态过程数据的分布因存在多个局部最大值,其密度曲线结构比单模态分布更为复杂。这使得传统 SPC 方法在处理多模态数据时可能表现不佳。在设计监控、监督和控制系統时,必须充分考虑过程数据的特性。然而,专门针对多模态数据设计的控制图在文献中有限。Quiñones-Grueiro 等人[2]对多模态过程的研究进行了总结,部分学者提出先通过转换函数将多模态特征去除,进而使用常规的监控方法解决问题。例如 Deng 等人[3]通过采用服从单峰分布的局部概率密度估计,将多模态数据转化为概率密度。然而,对于复杂系统而言,转换函数很难准确确定,并且数据转换的计算成本可能较高,为每个观测定义标准化模型是具有挑战性的。

另一种可行的选择是非参数或无分布假设的控制图,这类方法虽不依赖于某一具体的参数分布形式,但是,这些非参数控制图通常依赖于过程是单模态的假设,即数据是源自相同分布,因此不适用于存在多模态过程的模式例如, Hackl 和 Ledolter [4]提出的单个观测值的指数加权移动平均(EWMA)图是由观测值的“标准化秩”构建的,这是由受控(In-Control, IC)分布决定的。若分布未知,他们建议使用收集到的参考数据中的排名代替。该方法侧重于检测位置参数的变化,但是在监测整个多模态数据过程分布是否发生偏移时,其余的分布特征,如尺度参数,局部最大值的个数也是重要的质量指标。Qiu 和 Li [5]提出的对观测数据进行分类,然后利用一定的分类数据分析统计程序构造非参数 SPC 图。这种基于数据分类的方法会因为分类数量较多时容易导致列联表稀疏,监控方法在一定程度上会造成信息丢失。鉴于上述问题,针对单变量多模态过程开发既稳健又具有普适性的质量监控方法具有一定的研究意义。

本文提出了一种基于密度估计的数据驱动的监测方案,用于单变量多模态数据的在线监测。密度估

计已广泛应用于各种应用中。通过估计给定数据集的概率密度函数(PDF),可以了解数据的基本分布。基于 GMM 的方法对多模态过程进行建模,并且考虑了不同的模型估计方法[6][7]。本文针对单变量多模态过程的质量监控问题,提出了一种基于高斯混合模型的负对数似然值的 EWMA 监控方案。这种数据驱动的监控方法,能够有效地检测模型的参数偏移和模型的组分变化。本文的其余部分组织如下。第 2 节介绍了数据驱动的方法,包括高斯混合模型,模型估计的方法和监控方案的构建。第 3 节通过蒙特卡洛模拟对所提出的监测方案在不同失控情景下进行了比较分析。在第 4 节中讨论了在实际案例中所提出的监控方案的有效性。最后对论文进行总结并给出了未来的可能的研究方向。

## 2. 方法

假设  $X = \{x_1, x_2, \dots, x_t\}$  为一组单变量多模态数据,  $x_t$  是在  $t$  时刻的观察值。在对多模态数据的第二阶段监测过程中,为了监测一个连续的一元多模态过程  $\{x_t, t \geq 1\}$ , 假设存在一个 IC 数据集  $\{x_1, x_2, \dots, x_n\}$ , 其样本量为  $n$ , 预先从 IC 分布  $f_0$  中采集。我们的目的是确定过程是否已从 IC 状态变为失控(Out-of-Control, OC)状态,即可表示为假设检验:

$H_0: \{x_1, x_2, \dots, x_t\}$  遵循受控分布  $f_0$ ;

$H_1: \exists \tau \in [1, t], \{x_1, x_2, \dots, x_\tau\}$  遵循受控分布  $f_0$ ,  $\{x_{\tau+1}, \dots, x_t\}$  遵循失控分布  $f_1$ 。

其中  $\tau$  为变化点,  $f_1$  为失控分布, 满足  $f_0 \neq f_1$ 。对于含有多个峰的多模态过程[7], 指出其备择假设  $H_1$  可能存在两种类型的过程 OC 情景, 第一类是关于分布中已有峰的参数发生偏移, 第二类是关于分布的峰数发生变化。因此, 在多模态过程质量监控方法应该具有可以及时准确地监测出这两种 OC 类型的能力。

本文提出的监测单变量多模态数据的方案包括以下三个步骤:

步骤 1: 使用高斯混合模型方法对多模态过程进行建模。

步骤 2: 讨论不同的参数估计方法, 并进行比较分析。

步骤 3: 根据步骤 2 中得到的最优估计值, 构建 EWMA 的统计量, 并进行在线监测。

### 2.1. 高斯混合模型

高斯混合模型(Gaussian Mixture Model, GMM)是一种概率密度估计方法, 它假设数据由若干个高斯分布混合生成。考虑数据集  $X = \{x_1, x_2, \dots, x_n\}$ , 每个数据点都假定是由  $K$  个高斯分布的混合而成。高斯混合模型定义为:

$$f(x) = \sum_{j=1}^K w_j \mathcal{N}(x | \theta_j) \quad (1)$$

其中,  $w_j$  为第  $j$  个高斯分量的权重, 满足  $\sum_{j=1}^K w_j = 1$  且  $0 \leq w_j \leq 1$ ;  $\theta_j = \{\mu_j, \sigma_j^2\}$  为第  $j$  个高斯分量的分布参数;  $\mathcal{N}(x | \theta_j)$  是高斯分布, 表示对于给定的  $x$ , 其均值为  $\mu_j$ , 方差为  $\sigma_j^2$ 。

### 2.2. 模型估计的方法

GMM 中的关键问题是如何选择适当的分量数  $K$ 。常用的模型评估准则包括赤池信息量准则(Akaike Information Criterion, AIC) [8]和贝叶斯信息量准则(Bayesian Information Criterion, BIC) [9]:

$$AIC = 2k - 2 \ln(L) \quad (2)$$

$$BIC = \ln(N)k - 2 \ln(L) \quad (3)$$

其中,  $L$  为模型的最大似然估计;  $N$  为样本的数量;  $k$  为模型中自由参数的数量(GMM 中包括了每个高斯分布的均值、方差和权重), 对于一维 GMM, 每个高斯成分有两个参数  $\mu_j, \sigma_j^2$ , 且每个成分的权重  $w_j$  有  $K-1$  个自由参数, 因此自由参数的总数量为:

$$k = K(2+1) - 1 = 3K - 1 \quad (4)$$

AIC 和 BIC 通过惩罚复杂模型避免过拟合, 选择对应最小值的  $K$  作为 GMM 最优解, 分别记为 AIC-GMM 和 BIG-GMM。在 AIG-GMM 和 BIG-GMM 模型中, 采用期望最大化(Expectation-Maximization, EM) 算法[10]估计 GMM 的参数:

E 步: 计算分模型  $j$  对观测数据  $x_i$  的响应度  $\gamma_{ij}$ :

$$\gamma_{ij} = \frac{w_j \mathcal{N}(x_i | \theta_j)}{\sum_{j=1}^K w_j \mathcal{N}(x_i | \theta_j)}, i = 1, 2, \dots, n, j = 1, 2, \dots, K \quad (5)$$

M 步: 计算新一轮迭代的模型参数:

$$\hat{\mu}_j = \frac{\sum_{i=1}^n \hat{\gamma}_{ij} x_i}{\sum_{i=1}^n \hat{\gamma}_{ij}}, j = 1, 2, \dots, K \quad (6)$$

$$\hat{\sigma}_j^2 = \frac{\sum_{i=1}^n \hat{\gamma}_{ij} (x_i - \mu_j)^2}{\sum_{i=1}^n \hat{\gamma}_{ij}}, j = 1, 2, \dots, K \quad (7)$$

$$\hat{\alpha}_j = \frac{n_j}{n} = \frac{\sum_{i=1}^n \hat{\gamma}_{ij}}{n}, j = 1, 2, \dots, K \quad (8)$$

最终迭代计算到参数收敛。

AIC 和 BIC 需预设分量数  $K$ 。在数据量大且符合混合高斯假设时, 二者可以有效估计真实分量数。但对于非高斯数据, 可能导致组分接近、产生冗余簇。为此, 我们考虑了变分贝叶斯狄利克雷过程高斯混合模型(Variational Bayesian Dirichlet Process Gaussian Mixture Model, VBDP-GMM) [6] [11]。VBDP-GMM 引入狄利克雷过程(Dirichlet Process, DP)作为先验来自动选择模型的分量数  $K$ 。DP 定义为  $G \sim DP(\alpha, G_0)$ , 其中,  $\alpha$  为浓度参数, 控制新类别的产生频率,  $G_0$  为基准分布(通常为高斯分布)。在 VBDP-GMM 模型中, 通过变分推断来估计模型的参数  $\theta = \{\mu_j, \sigma_j^2, w_j, z_i\}$  (各高斯分量的均值  $\mu_j$ , 方差  $\sigma_j^2$ 、权重  $w_j$  以及隐藏变量(数据点  $x_i$  的分配分量  $z_i$ )。在变分推断中, 找到一个变分分布  $q(\theta)$  来近似真实的后验分布  $p(\theta|X)$ , 变分目标是 minimize 变分分布  $q(\theta)$  与真实后验分布  $p(\theta|X)$  的 KL 散度:

$$KL(q\|p) = \int q(\theta) \log \frac{q(\theta)}{p(\theta|X)} d\theta \quad (9)$$

由于  $p(\theta|X)$  包含棘手的边缘似然  $p(X)$ , 因此直接 KL 最小化是不可行的。经过式(9)的数学变形, 我们得到以下结果:

$$\begin{aligned} L(q) &= E_{q(\theta)} \left[ \log \frac{p(X, \theta)}{q(\theta)} \right] \\ &= E_{q(\theta)} \left[ \log \frac{p(\theta) p(X|\theta)}{q(\theta)} \right] \\ &= E_{q(\theta)} \left[ \log p(X|\theta) \right] - KL(q(\theta)\|p(\theta)) \end{aligned} \quad (10)$$

因此, 最小化 KL 散度等价于最大化变分下界(Evidence Lower Bound, ELBO)。通过优化 ELBO, 我们可以得到模型的近似后验分布, 从而估计模型参数。

### 2.3. 监控方案

在本小节中, 提出了基于 GMM 的阶段 II 的监测方案, 以实现单变量多模态过程的 OC 状态的有效检测。在阶段 II 的过程监控中, 我们目标是检测过程是否从 IC 状态偏移到 OC 状态。由于多模态数据呈现复杂的分布形态(如多峰分布), 直接使用传统的参数化控制图(如 Shewhart 图、CUSUM 或 EWMA)可能无法有效检测其变化。根据假设检验中的备择假设  $H_1$  表明该过程可能存在两种类型的过程 OC 情景, 关于模型的参数发生偏移以及模型的峰数发生变化。因此, 在多模态过程质量监控方法应该具有可以及时准确地监测出这两种 OC 类型的能力。

GMM 通过对历史正常数据集进行训练, 为每个新的观测值  $x_i$  提供密度估计  $\hat{f}(x_i)$ , 并通过计算对应的对数似然值(Log-Likelihood, LL)来监控过程中的异常变化。当观测值偏离 IC 状态时, 其在当前密度估计模型中的出现概率显著下降, 导致对应的 LL 值减小。为了在多模态过程中实现有效的在线监测, 我们将计算新观测值的负对数似然值(NLL)统计量[12], 并与控制限比较, 判断是否出现 OC。对于第  $i$  个观测值的 NLL 值统计量定义如下:

$$y_i = -\mathcal{L}_i = -\log\left(\sum_{j=1}^K \hat{w}_j \mathcal{N}(x|\hat{\theta}_j)\right) \quad (11)$$

其中,  $\mathcal{L}_i = \log(\hat{f}(x_i))$  是第  $i$  个观测值的 LL 值。

为了适应多模态数据过程建模并进行公平比较, 本文构建了基于公式(11)中负对数似然值(NLL)的非参数指数加权移动平均(EWMA)统计量

$$Q_i = \lambda y_i + (1-\lambda)Q_{i-1}, i=1,2,\dots,n \quad (12)$$

其中,  $E_0$  是初始值, 本文设置初始值  $E_0 = 0$ 。  $\lambda$  是平滑系数, 一般来说对于 EWMA 类型的控制图, 选择一个相对较小的  $\lambda$  可以更好地检测到小的偏移, 反之较大的  $\lambda$  则对较大的偏移更为敏感[13]。这一原理同样适用于本文控制图, 为了进行比较研究, 本文选择了  $\lambda \in \{0.05, 0.1, 0.2\}$  三个值。控制线  $H$  是一个关键阈值, 用于判断多模态数据是否偏离了正常状态。我们可以采用二分搜索算法来计算 IC 数据的控制线  $H$  值。

## 3. 数值性能比较

### 3.1. 主要竞争方法

第二个对比方法我们考虑的是将无分布控制图或非参数控制图作为比较方法。Hackl 和 Ledolter 提出了基于“标准秩”的 EWMA 图表(简称 HLE) [4], 这些标准秩是使用受控分布作为历史参考样本计算的。也就是说, 在受控分布未知的情况下, 可以从受控过程中选取一个大小为  $n$  的历史样本  $\{x_1, x_2, \dots, x_n\}$  作为参考, 并定义新观测值  $x_i$  的标准化秩为

$$R_i = \frac{2}{n+1} \left( R_i^* - \frac{n}{2} \right) \quad (13)$$

其中,  $R_i^*$  是  $x_i$  关于  $n$  个历史样本  $\{x_1, x_2, \dots, x_n\}$  的秩, 即

$$R_i^* = 1 + \sum_{i=1}^n I(x_i \leq x_i) \quad (14)$$

其中,  $I(x_i \leq x_t)$  是指示性函数, 表示当  $x_i \leq x_t$  时,  $I(x_i \leq x_t) = 1$ , 否则为 0。于是 HLE 控制图的统计量为

$$T_t = \lambda R_t + (1 - \lambda) T_{t-1}, t = 1, 2, \dots, n \quad (15)$$

其中,  $T_0 = 0$  且  $\lambda$  是一个常数。当  $T_t \geq H$  时, 则发出失控警报,  $H$  是由  $ARL_0$  得到的控制线。

### 3.2. 控制图参数设置(模型估计的性能)

#### (1) 多模态模型的设置

在仿真实验中, 首先需要生成服从单变量多模态过程的 IC 数据集。我们基于不同的分布类型的混合, 构建了多模态过程分布模型。本文考虑了两种不同的模型, 定义如下:

$$\text{Model A } f = 0.5p_1(x) + 0.5p_2(x)$$

$$\text{Model B } f = 0.2p_1(x) + 0.2p_2(x) + 0.2p_3(x) + 0.2p_4(x) + 0.2p_5(x)$$

其中,  $p_j(x)$  是样本  $x$  的在第  $j$  簇的密度函数。对于多模态过程分布的每个峰, 考虑了非高斯分布(均匀分布)以及高斯分布进行模拟分析, 模型参数设置如表 1 所示。

**Table 1.** Model parameter settings

**表 1.** 模型参数设置

分布类型	模型	$p_1(x)$	$p_2(x)$	$p_3(x)$	$p_4(x)$	$p_5(x)$
均匀分布 $U(a,b)$	A-1	$U(0,2)$	$U(5,7)$			
	B-1	$U(0,1)$	$U(3,5)$	$U(6,7)$	$U(9,11)$	$U(12,15)$
高斯分布 $N(\mu, \sigma^2)$	A-2	$N(0,1)$	$N(0,3)$			

#### (2) 控制图参数设置和评价指标

在所有图表中, 我们设置历史数据集个数  $n = 300$ ,  $ARL_0 = 200$ 。平滑系数  $\lambda \in \{0.05, 0.1, 0.2\}$ , 初始值  $E_0$  设为 0。通过 10,000 次重复仿真, 采用二分搜索算法, 逼近不同控制图的控制限。控制图的监控性能评价指标, 本文主要分析了控制图的 OC 性能指标——OC 平均运行长度( $ARL_1$ )和总体性能指标——相对平均指数 RMI [14]。其中 RMI 定义如下:

$$RMI = \frac{1}{m} \sum_{i=1}^m \frac{ARL_{\delta_i} - \min ARL_{\delta_i}}{\min ARL_{\delta_i}}. \quad (16)$$

其中,  $m$  是所考虑的偏移数量,  $ARL_{\delta_i}$  是给定控制图在偏移为  $\delta_i$  时的  $ARL_1$  值, 而  $\min ARL_{\delta_i}$  是所有控制图在偏移为  $\delta_i$  时的最小的  $ARL_1$  值。在相同的偏移情景和相等的  $ARL_0$  值下, 具有较小  $ARL_1$  的控制图表现出更佳的性能。从整体性能指标的角度来看, 在给定的偏移范围内, 具有最小 RMI 的控制图性能最优。

### 3.3. 数值模拟比较分析

#### 3.3.1. GMM 的模型估计效果的比较结果

在对模型的监控性能比较之前, 我们首先对 GMM 的估计性能进行了对比分析。具体而言, 当模型的真实分量数未知时, 对于 AIC-GMM 和 BIC-GMM 方案, 我们预先设定初始分量数  $K$  的范围为 2 到 40, 而 VBDP-GMM 利用 DP 自适应地选择模型中的分量数无需预设分量数。然后对表 1 中给出的方案进行了估计, 最终算法收敛时的分量数如表 2 所示。

由表 2 可知, 在估计由高斯分布生成的多模态数据时, 三种方案都收敛到了真实分量数。然而, 上述三种方案都假设每个组分均由高斯分布生成的, 因此它们在捕捉高斯分布混合的多模态数据时表现较好。然而, 在拟合非高斯数据时, 均出现了不同程度的偏差。其中 AIC-GMM 与 BIC-GMM 出现的过拟合尤为严重。由于无法准确估计分量数可能会导致 GMM 不能有效地描述这些复杂数据, 于是在后续讨论中, 将不再考虑。基于 VBDP-GMM 的 NLL 值的 EWMA 监控方案, 记为 VGNE。

对于模拟, 为了验证本文所提出的控制图的有效性, 我们考虑了基于非高斯和高斯多模态过程的两类具有代表性的 OC 情景: 多模态过程中的参数偏移, 包括位置参数和尺度参数; 多模态过程中的分量数变化。

**Table 2.** Number of components at convergence for different algorithms

**表 2.** 不同算法收敛时的分量数

序号	真实分量数	AIC-GMM	BIC-GMM	VBDP-GMM
A-1	$K=2$	$K=8$	$K=4$	$K=2$
B-1	$K=5$	$K=18$	$K=9$	$K=5$
A-2	$K=2$	$K=2$	$K=2$	$K=2$

### 3.3.2. 非高斯多模态过程偏移时的性能

对于模型 A-1, OC 情景我们分别考虑了: (a) 位置参数均偏移  $U(a,b)+\delta$ ,  $\delta \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.75, 1, 2\}$ , 以模拟过程发生不同幅度的位置偏移; (b) 为了保证在尺度参数发生偏移时, 位置参数保持不变, 我们设定尺度参数均偏移的失控情况为  $\delta_1 \cdot \left( U(a,b) - \frac{a+b}{2} \right) + \frac{a+b}{2}$ ,  $\delta_1 \in \{1.1, 1.2, 1.3, 1.4, 1.5, 1.75, 2, 3\}$ , 表示尺度参数的缩放系数, 对应过程标准差放大至原来的 1.1 倍、1.2 倍等; (c) 添加新的组分。表 3 给出了多个分量的位置参数偏移的性能比较结果; 表 4 给出了多个分量的尺度参数偏移下控制图的性能比较结果; 表 5 给出了对于模型分量数变化的 OC 示例, 表 6 给出了模型分量数变化 OC 示例下控制图的性能比较结果。从上述表中可得, 无论 OC 情景是模型参数偏移还是模型组分数变化, VGNE 控制图的 ARL1 和 RMI 值都达到最低, 都可以针对二阶的非高斯多模态过程进行有效且稳健的在线监控。

**Table 3.** Performance of control charts under location parameter shifts in multiple components

**表 3.** 对于多个分量的位置参数偏移下控制图的性能比较

$\delta$	VGNE			HLE		
	$\lambda = 0.05$	$\lambda = 0.1$	$\lambda = 0.2$	$\lambda = 0.05$	$\lambda = 0.1$	$\lambda = 0.2$
0	200.99	199.60	200.48	200.06	200.66	200.63
0.1	149.12	124.36	<b>105.85</b>	95.00	103.22	107.30
0.2	96.49	67.55	<b>47.75</b>	55.46	60.07	65.64
0.3	68.83	42.86	<b>27.60</b>	37.81	40.92	43.85
0.4	52.06	31.08	<b>18.86</b>	28.23	29.94	32.89
0.5	41.65	23.99	<b>14.07</b>	22.34	23.91	25.66
0.75	26.47	14.70	<b>8.43</b>	15.20	15.34	15.91
1	18.28	10.09	<b>5.75</b>	11.92	11.77	11.65

续表

2	6.90	3.86	<b>2.28</b>	8.45	8.49	7.77
RMI	0.521	0.195	<b>0.0086</b>	0.214	0.239	0.257
H	1.670	1.757	1.882	0.165	0.275	0.432

**Table 4.** Performance of control charts under scale parameter shifts in multiple components**表 4.** 对于多个分量的尺度参数偏移下控制图的性能比较

$\delta_1$	VGNE			HLE		
	$\lambda = 0.05$	$\lambda = 0.1$	$\lambda = 0.2$	$\lambda = 0.05$	$\lambda = 0.1$	$\lambda = 0.2$
1	200.99	199.60	200.48	200.06	200.66	200.63
1.1	82.57	58.57	<b>44.40</b>	186.38	183.07	170.92
1.2	56.82	35.40	<b>23.73</b>	178.64	169.30	152.84
1.3	44.86	26.30	<b>16.41</b>	170.28	156.17	137.24
1.4	37.44	21.40	<b>12.84</b>	161.15	148.13	130.32
1.5	32.22	18.05	<b>10.65</b>	155.73	142.56	122.41
1.75	23.70	13.15	<b>7.52</b>	146.95	131.54	107.86
2	18.52	10.25	<b>5.87</b>	141.71	124.68	102.79
3	9.70	5.49	<b>3.25</b>	130.34	108.65	85.26
RMI	0.123	0.043	<b>0.001</b>	0.808	0.729	0.618
H	1.670	1.757	1.882	0.165	0.275	0.432

**Table 5.** OC case for varying number of model components**表 5.** 模型分量数变化的 OC 示例

OC case	$w_j$	模型分量数变化
1-1	0.3, 0.5, 0.2	$U(0,2), U(5,7), U(0.5,2.5)$
1-2	0.3, 0.5, 0.2	$U(0,2), U(5,7), U(1,3)$
1-3	0.3, 0.5, 0.2	$U(0,2), U(5,7), U(2,4)$
2-1	0.5, 0.3, 0.2	$U(0,2), U(5,7), U(5.5,7.5)$
2-2	0.5, 0.3, 0.2	$U(0,2), U(5,7), U(6,8)$
2-3	0.5, 0.3, 0.2	$U(0,2), U(5,7), U(7,9)$

**Table 6.** Performance of control charts under the OC case with varying numbers of model components**表 6.** 对于模型分量数变化 OC 示例下控制图的性能比较

OC case	VGNE			HLE		
	$\lambda = 0.05$	$\lambda = 0.1$	$\lambda = 0.2$	$\lambda = 0.05$	$\lambda = 0.1$	$\lambda = 0.2$
1-1	97.80	67.25	<b>47.13</b>	117.25	134.42	159.13
1-2	50.55	30.38	<b>19.45</b>	81.82	102.75	129.81
1-3	25.70	14.60	<b>8.88</b>	62.02	79.62	112.18
2-1	84.92	56.23	<b>37.85</b>	88.37	87.72	86.06
2-2	46.70	27.63	<b>17.63</b>	57.36	56.11	51.54

续表

2-3	22.08	12.79	<b>7.93</b>	43.92	41.75	37.42
RMI	0.345	0.126	<b>0.000</b>	0.580	0.638	0.714
H	1.670	1.757	1.882	0.165	0.275	0.432

除了考虑二阶的情况，本文也考虑了基于均匀分布五阶的多模态过程模型 B-1，进一步说明本文所提出的 VGNE 控制图在多模态过程的监控性能。OC 情景仍然考虑模型参数偏移和模型分量数变化，失控情况分别设置为：(a) 位置参数均偏移  $U(a,b)+\delta$ ，(b) 尺度参数均偏移  $\delta \cdot \left( U(a,b) - \frac{a+b}{2} \right) + \frac{a+b}{2}$ ，(c) 添加新的组分  $U(12,15)+\delta$ 。由表 7 给出了基于非高斯多模态过程模型 B-1 下控制图的监控性能结果，在位置参数偏移较小的情况下，HLE 图监控性能优于 VGNE 图，这是因为 HLE 图的设计针对位置参数变化更为敏感。但当偏移量增大时，VGNE 图的监控性能更优，并且在尺度参数发生偏移以及注入新的组分时，VGNE 控制图都可以针对五阶的非高斯多模态过程进行有效且稳健的在线监控。

**Table 7.** Monitoring performance of control charts under the non-Gaussian multimodal process model B-1

**表 7.** 基于非高斯多模态过程模型 B-1 下控制图的监控性能

OC case	$\delta$	VGNE			HLE			
		$\lambda = 0.05$	$\lambda = 0.1$	$\lambda = 0.2$	$\lambda = 0.05$	$\lambda = 0.1$	$\lambda = 0.2$	
位置参数	a	0.2	184.01	167.54	138.08	<b>96.33</b>	105.21	118.22
		0.5	121.73	91.23	67.41	<b>50.54</b>	58.76	69.11
		1	70.80	42.97	<b>28.15</b>	32.45	35.76	41.31
尺度参数	b	1.2	142.86	118.63	<b>96.39</b>	189.74	184.46	187.62
		1.5	107.17	75.40	<b>53.30</b>	186.81	183.81	175.91
		2	82.27	51.32	<b>33.60</b>	183.76	183.70	173.82
组分数	c	0.5	136.67	110.15	<b>89.12</b>	170.37	162.31	159.45
		1	104.83	72.65	<b>50.85</b>	147.97	143.74	136.76
		2	73.50	45.07	<b>27.98</b>	127.27	123.92	112.36
H		2.840	2.923	3.041	0.165	0.275	0.432	

### 3.3.3. 高斯多模态过程偏移时的性能

在本节中我们比较了 VGNE 和 HLE 控制图在高斯多模态过程中的监控性能。以表 1 的模型 A-2 为例，分析了模型参数偏移和模型分量数变化，OC 情况分别设置为：(a) 位置参数整体偏移  $N(\mu+\delta, \sigma^2)$ ，(b) 尺度参数整体偏移  $N(\mu, \delta^2)$ ，(c) 新增组分  $N(3, \delta^2)$ 。表 8 给出具有代表性的结果。

**Table 8.** Monitoring performance of control charts under the Gaussian multimodal process model A-2

**表 8.** 基于高斯多模态过程模型 A-2 下控制图的监控性能

OC case		VGNE			HLE			
		$\lambda = 0.05$	$\lambda = 0.1$	$\lambda = 0.2$	$\lambda = 0.05$	$\lambda = 0.1$	$\lambda = 0.2$	
位置参数	a	0.2	175.24	169.39	165.52	<b>81.19</b>	89.26	95.74
		0.5	108.59	93.49	90.41	<b>33.59</b>	36.31	41.06
		1	54.50	37.97	31.20	<b>15.76</b>	15.63	16.83

续表

		1.2	120.33	106.75	<b>101.35</b>	199.51	188.86	182.48
尺度参数	b	1.5	80.83	62.94	<b>54.98</b>	184.53	171.85	163.73
		2	56.44	38.82	<b>31.09</b>	167.22	153.37	137.51
		1.2	177.58	174.46	<b>165.99</b>	209.12	199.89	199.10
组分数	c	1.5	149.85	139.58	<b>131.20</b>	203.20	201.34	194.02
		2	120.88	103.32	<b>94.07</b>	213.08	199.75	186.10
		H	2.102	2.283	2.595	0.165	0.275	0.432

由表 8 可知, 由于 HLE 图被设计成只对位置变化敏感, HLE 图在检测位置参数发生偏移时更有效。但在尺度参数发生偏移时, 由于 VGNE 图善于捕捉高斯混合分布的特征, 在检测尺度参数中小偏移方面表现优秀的检测性能, 而 HLE 图的  $ARL_1$  却是 VGNE 控制图的几倍。在组分发生变化时, VGNE 图有更低的  $ARL_1$ , 而 HLE 图的  $ARL_1$  在预设的  $ARL_0 = 200$  附近变化, 难以检测到组分的变化。因此, VGNE 图能够对高斯多模态过程进行有效且稳健的在线监控。

综合上述分析, 本文系统评估了所提出的 VGNE 控制图与基准方法 HLE 控制图在多个候选平滑参数  $\lambda \in \{0.05, 0.1, 0.2\}$  下的性能表现。结果表明, 在所有失控场景下, 当  $\lambda = 0.2$  时, VGNE 控制图均取得了最优的 RMI 值。因此, 本文选取  $\lambda = 0.2$  作为最优平滑参数, 该取值在保持对小且持续偏移的敏感性与快速检测较大幅度偏移的能力之间取得了最佳平衡。后续的实例研究均基于该参数设置展开。

#### 4. 实例分析

半导体是一种具有介于导体和绝缘体之间的电导特性的材料。主要应用在集成电路、通信系统、光伏发电、大功率电源转换等领域。在统计过程控制中, 监测半导体制造过程主要通过收集和分析生产过程中的数据来实现。本节给出了一个实际数据示例, 以说明 VGNE 图的有效性。本文选用加州大学欧文分校的机器学习库中维护了一段从 2008 年 7 月至 10 月期间由自动管理半导体制造过程的计算机系统收集的半导体制造过程数据, 其原始数据可以在 UC Irvine Machine Learning Repository (<http://archive.ics.uci.edu/ml/datasets/SECOM>)网站上找到。该组半导体制造过程数据共包含 1567 个观测数据, 每个观测数据有 590 个质量特性, 其中前 1463 个观测数据为合格品, 剩余 104 个为不合格品。从各种属性中选择一个变量, 记作“X528”。考虑到数据量大, 去除异常值后, 我们将前 600 个数据作为训练样本, 后 104 个不合格数据作为测试样本。Shapiro-Wilk 正态性检验结果显示, X528 的 IC 样本的 p 值为  $7.20 \times 10^{-11}$ , 拒绝原假设, 表明数据显著偏离高斯分布。

由图 1 可知, 由测试样本得到的经验概率密度图存在两个峰值, 这表示由这 600 个数据作为 IC 数据集并非都来自同一个分布, 而是由两种分布不同的过程混合而成。因此这个半导体制造过程可能是多种操作模式的组合, 并且过程数据可能遵循多模态分布。当  $ARL_0 = 200$ ,  $\lambda = 0.2$  时, 绘制 VGNE 以及 HLE 控制图。

由图 2 可知, 对于后 104 个测试样本, VGNE 图偏移发生的第一个信号出现在第 7 次观测, HLE 图未发出报警信号, 结果显示我们的方案略优于 HLE。在实际应用中, 当控制图发出第一个信号时, 我们认为过程已经失控了, 因为在此信号之后所有的统计数据几乎都超过了控制线。此时, 操作员应立即即检查工艺设置和设备校准情况, 努力将工艺恢复到 IC 状态。如果初步调整未能解决问题, 则需要及时联系工艺工程师, 以进一步分析和解决潜在的工艺问题。

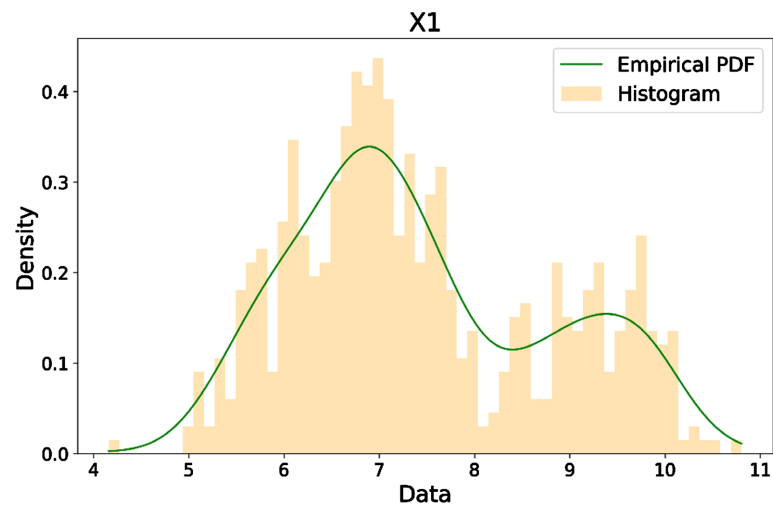


Figure 1. Empirical PDF of the IC dataset

图 1. IC 数据集的经验 PDF

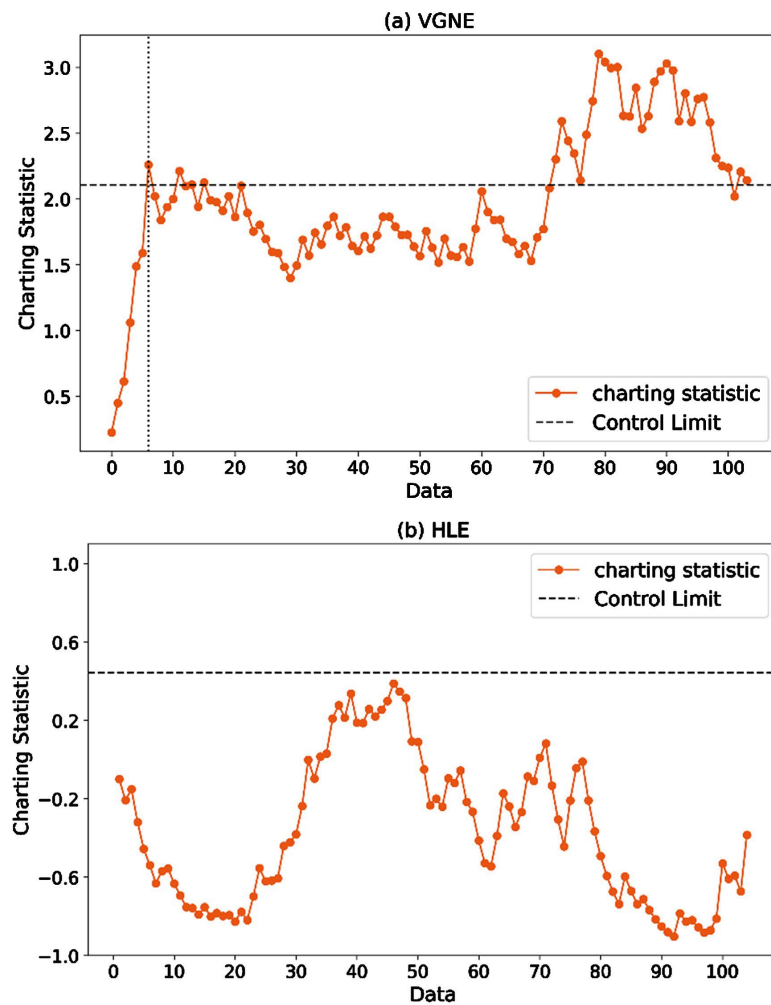


Figure 2. Monitoring of Phase II observations using VGNE and HLE control charts

图 2. VGNE 和 HLE 控制图用于阶段 II 观测数据的监控

## 5. 总结

多模态数据在许多复杂应用中都有出现，这类数据不满足单模态和高斯分布的基本假设，因此难以使用某一具体的参数分布对其进行建模。传统的统计过程控制方法在处理多模态数据时可能表现不佳。为解决这一限制，拟议的 VGNE 图表采用了数据驱动的方法，利用 DP 自适应地选择模型中的分量数无需预设分量数，该方法有效捕捉了数据固有的多模态性，展示了在高斯和非高斯多模态过程中稳健的监控性能。

最后，尽管我们提出的方法为多模态过程设计得复杂，VGNE 图表展现了显著的鲁棒性和广泛的适应性。其对未知分布的过程进行有效监控，而无需事先了解流程知识。目前的主要限制在于当 IC 观测不足时，在线监测的实施。因此，未来的改进将聚焦于多元扩展和自启动能力开发，以促进更灵活的实时应用。

## 参考文献

- [1] Montgomery, D.C. (2020) Introduction to Statistical Quality Control. John Wiley & Sons.
- [2] Quiñones-Grueiro, M., Prieto-Moreno, A., Verde, C. and Llanes-Santiago, O. (2019) Data-Driven Monitoring of Multimode Continuous Processes: A Review. *Chemometrics and Intelligent Laboratory Systems*, **189**, 56-71. <https://doi.org/10.1016/j.chemolab.2019.03.012>
- [3] Deng, X., Zhong, N. and Wang, L. (2017) Nonlinear Multimode Industrial Process Fault Detection Using Modified Kernel Principal Component Analysis. *IEEE Access*, **5**, 23121-23132. <https://doi.org/10.1109/access.2017.2764518>
- [4] Hackl, P. and Ledolter, J. (1991) A Control Chart Based on Ranks. *Journal of Quality Technology*, **23**, 117-124. <https://doi.org/10.1080/00224065.1991.11979302>
- [5] Qiu, P. and Li, Z. (2011) On Nonparametric Statistical Process Control of Univariate Processes. *Technometrics*, **53**, 390-405. <https://doi.org/10.1198/tech.2011.10005>
- [6] Blei, D.M. and Jordan, M.I. (2006) Variational Inference for Dirichlet Process Mixtures. *Bayesian Analysis*, **1**, 121-144. <https://doi.org/10.1214/06-ba104>
- [7] Wang, K., Li, J. and Tsung, F. (2018) Adaptive Monitoring of Multimodal Data. *Computers & Industrial Engineering*, **125**, 364-374. <https://doi.org/10.1016/j.cie.2018.09.001>
- [8] Akaike, H. (1974) A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, **19**, 716-723. <https://doi.org/10.1109/tac.1974.1100705>
- [9] Schwarz, G. (1978) Estimating the Dimension of a Model. *The Annals of Statistics*, **6**, 461-464. <https://doi.org/10.1214/aos/1176344136>
- [10] Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977) Maximum Likelihood from Incomplete Data via the *em* Algorithm. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **39**, 1-22. <https://doi.org/10.1111/j.2517-6161.1977.tb01600.x>
- [11] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O. and Duchesnay, É. (2011) Scikit-Learn: Machine Learning in Python. *The Journal of Machine Learning Research*, **12**, 2825-2830.
- [12] Wang, Z., Gong, R., Song, L., He, S. and Gao, Y. (2024) A Data-Driven Monitoring Scheme for Multivariate Multimodal Data. *Computers & Industrial Engineering*, **192**, Article ID: 110186. <https://doi.org/10.1016/j.cie.2024.110186>
- [13] Lucas, J.M. and Saccucci, M.S. (1990) Exponentially Weighted Moving Average Control Schemes: Properties and Enhancements. *Technometrics*, **32**, 1-12. <https://doi.org/10.1080/00401706.1990.10484583>
- [14] Han, D. and Tsung, F. (2006) A Reference-Free Cuscore Chart for Dynamic Mean Change Detection and a Unified Framework for Charting Performance Comparison. *Journal of the American Statistical Association*, **101**, 368-386. <https://doi.org/10.1198/016214505000000556>