Statistics and Application 统计学与应用, 2026, 15(2), 193-199
Published Online February 2026 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2026.152047

Hans汉斯

基于经验似然比检验对多数据流的在线监控

莫文玉，齐德全*

长春理工大学数学与统计学院，吉林 长春

收稿日期：2026年1月26日；录用日期：2026年2月17日；发布日期：2026年2月27日

摘　要

随着人工智能与大数据的飞速发展，对多数据流的实时在线监控已成为智能制造与质量管理的核心需求。从统计过程控制的角度，提出了监控复杂结构的多数据流均值是否发生漂移的变点模型。鉴于部分数据流的偏态或长尾等特点，将单个数据流的经验似然比检验统计量转化为Q统计量。为了同时监控多数据流的中小漂移，通过Q统计量建立Max-EWMA控制图进行在线监控。以威布尔分布、指数分布、对数正态分布和t分布为例，通过蒙特卡洛模拟研究所给出的在线监控方法的性能。模拟结果表明，该控制图对监控中小漂移具有较理想的性能。

关键词

经验似然比检验，控制图，数据流，统计过程控制

Online Monitoring of Multiple Data Streams Based on Empirical Likelihood Ratio Test

Wenyu Mo, Dequan Qi*

School of Mathematics and Statistics, Changchun University of Science and Technology, Changchun Jilin

Received: January 26, 2026; accepted: February 17, 2026; published: February 27, 2026

Abstract

With the rapid development of artificial intelligence and big data, real-time online monitoring of multiple data streams has become a core requirement for intelligent manufacturing and quality management. From the perspective of statistical process control, a change-point model is proposed to monitor whether the mean of complex-structured multi-data streams experiences drift. Considering characteristics such as skewness or long-tailed distributions in certain data streams, the empirical likelihood ratio test statistic for individual data streams is transformed into a Q-statistic. To

*通讯作者。

文章引用: 莫文玉, 齐德全. 基于经验似然比检验对多数据流的在线监控[J]. 统计学与应用, 2026, 15(2): 193-199.
DOI: 10.12677/sa.2026.152047

simultaneously monitor small and medium drifts across multiple data streams, a Max-EWMA control chart is established using the *Q*-statistic for online monitoring. Taking the Weibull distribution, exponential distribution, log-normal distribution, and t-distribution as examples, the performance of the proposed online monitoring method is investigated through Monte Carlo simulations. The simulation results demonstrate that this control chart exhibits ideal performance in monitoring small and medium drifts.

Keywords

Empirical Likelihood Ratio Test, Control Chart, Data Stream, Statistical Process Control

1. 引言

在人工智能与大数据时代，交叉学科不断促进统计学的发展[1]。多数据流的相关问题一直吸引着学者们从不同的视角展开研究。Mei (2010)提出最大化所有时刻所有数据流的似然比统计量的方法，对多数据流进行监控[2]。Du 和 Zou (2018)基于指数加权滑动平均(Exponentially Weighted Moving Average, EWMA)控制图，给出了一个新的错误发现率控制方法，监控多数据流问题[3]。Ren 等(2020)研究了存在聚类或空间模式时数据流的监控问题，设计了一种定向抽样模式结合似然比检验进行自适应检测[4]。Dassanayake 和 French (2024)通过控制图的方法发现疾病的爆发区域[5]。

关于多数据流的监控，大多假设数据流是简单结构的，每个数据流都是一元的，且服从同一分布。而统计过程控制(Statistical Process Control, SPC)的理论与应用研究除了经典的适用于正态数据的控制图以外，成果也非常丰富。Zhou 等(2012)建立了加权似然比检验统计量监控变样本量的泊松数据[6]。Li 等(2014)研究了适用于多元二项及多元多项数据的控制图[7]。范倩(2018)建立了基于似然比检验的对数正态分布控制图[8]。郭宝才等(2018)设计了参数未知下基于定数截尾样本的控制图监控指数分布的数据[9]。曹程明(2019)开展了监控威布尔分布数据的控制图的研究[10]。Sanusi 等(2020)建立了 Max-EWMA 控制图用来监控伽马和指数分布的数据[11]。刘英杰(2022)开展了监控多元离散型数据的 CUSUM 控制图的研究[12]。廉惠然和齐德全(2023)研究了指数分布的订货周期数据等网络销售数据的监控方法[13]。

综上，鉴于数据分布可能存在的多样性，本文考虑以下复杂结构的多数据流的在线监控问题，仅给出监控均值或均值向量是否发生漂移的一般框架。单个数据流可以是一元的也可以是多元的，可以是连续型的随机变量也可以是离散型的随机变量。为提高对离散型数据、偏态或长尾等连续型数据的稳健性通过经验似然比检验(Empirical Likelihood Ratio Test, ELRT)建立监控统计量，而且 ELRT 既适用于单个一元的数据流，也适用于单个多元的数据流。为了更好地对多个数据流同时进行监控，将 ELRT 变换为 Q 统计量[14]。为了利用历史数据的信息，提高对中小漂移的监控效果，通过 Q 统计量建立 EWMA 型的控制图。在数据流之间相互独立的假设下，最后通过 Max-EWMA [11]构造监控统计量进行在线监控。蒙特卡洛模拟表明，所提出的方法对不同的漂移量都有较好的表现。

2. 变点模型

假设在智能制造或质量管理等过程中，需要监控 N 条相互独立的数据流 $\boldsymbol{X}_1,\cdots,\boldsymbol{X}_N$。第 n 条数据流根据问题背景用一元随机变量或多元随机向量 $\boldsymbol{X}_n$ 来刻画，其均值或均值向量记为 $\boldsymbol{\mu}_n$，其方差或协方差矩阵

记为$\boldsymbol{\Sigma}^{(n)}$，并假设在整个监控过程中$\boldsymbol{\Sigma}^{(n)}$保持不变。假设存在一个未知的时刻$\tau_n$，第$n$条数据流失控，即均值$\boldsymbol{\mu}_n$由可控时的$\boldsymbol{\mu}_n^0$变为失控时的$\boldsymbol{\mu}_n^1$。于是，在每一时刻$t=1,2,\cdots$监控以下变点模型：

$$H_{n,t}^0:\mu_{n,1}=\cdots=\mu_{n,t}=\mu_n^0,\ H_{n,t}^1:\mu_{n,1}=\cdots=\mu_{n,\tau_n}=\mu_n^0,\mu_{n,\tau_n+1}=\cdots=\mu_{n,t}=\mu_n^1,n=1,\cdots,N.$$

这里，可能有的数据流是一元的，有的数据流是多元的；可能有的数据流是离散型的，有的数据流是连续型的；可能有的数据流是对称分布的，有的数据流是偏态分布的；可能有的数据流是分布已知的，有的数据流是分布未知的。针对这样复杂结构的数据流监控问题，建议通过如下非参数方法进行在线监控。

3. 基于经验似然比检验的在线监控方法

在时刻t，对第n条数据流抽取m_n个样本$\left(X_{nt1},\cdots,X_{ntm_n}\right)$。鉴于部分数据流可能是偏态分布或分布未知，结合 ELRT 与 Max-EWMA 方法进行在线监控。

首先，对第n条数据流计算 ELRT 统计量，并利用其渐近性将其转化为Q统计量。为每个样本X_{nti}分配一个概率权重P_i，满足$P_i\geq 0$，且$\sum_{i=1}^{m_n}P_i=1$。经验似然函数定义为

$$L(F)=\prod_{i=1}^{m_n}P_i.$$

在原假设$H_{n,t}^0$下，根据约束条件$\sum_{i=1}^{m_n}P_iX_{nti}=\mu_n^0$，通过拉格朗日乘子法，最大化经验似然函数$L(F)$，得到一组最优权重$P_i^0$和对应的约束最大值$L\left(\mu_n^0\right)$。无约束的最大经验似然就是给每个样本赋予权重$1/m_n$，对应的经验似然值为$L\left(\bar{X}\right)=m_n^{-m_n}$。进一步得 ELRT 统计量为

$$R\left(\mu_n^0\right)=\frac{L\left(\mu_n^0\right)}{L\left(\bar{X}\right)}=\prod_{i=1}^{m_n}\left(m_nP_i^0\right)$$

在$H_{n,t}^0$成立且满足一些正则条件下，当$m_n\to\infty$时，统计量$-2\log\left(R\left(\mu_n^0\right)\right)$依分布收敛于自由度为$q$($q$是该数据流的维数)的卡方分布。于是构造如下$Q$统计量

$$Q_{nt}=\Phi^{-1}\left(H\left(-2\log\left(R\left(\mu_n^0\right);q\right)\right)\right)$$

其中，$\Phi^{-1}(\cdot)$是标准正态分布的分布函数的反函数，$H(\cdot;q)$是自由度为q的卡方分布的分布函数。

然后，计算 EWMA 序列

$$S_{1t}=(1-\lambda)S_{1t-1}+\lambda Q_{1t},\cdots,S_{Nt}=(1-\lambda)S_{Nt-1}+\lambda Q_{Nt},t=1,2,$$

其中初始值$S_{10}=\cdots=S_{N0}=0$，$\lambda\in(0,1]$为光滑参数。

最后，计算 Max-EWMA 统计量

$$M_t=\mathrm{Max}\left(S_{1t},\cdots,S_{Nt}\right).$$

当$M_t\geq h$时，发出过程失控的警报，其中$h>0$是控制线，满足可控时的平均运行长度为 ARL_0。

在线监控的流程图如图 1 所示。

4. 统计模拟

通过蒙特卡洛模拟验证所提出方法(简记为 Max-EWMA)监控复杂结构数据流的有效性，将N个数据流的Q统计量取最大作为对比方案(简记为 Max-Q)。进行 1000 次重复模拟实验，调整不同控制图方法的控制线使得可控时的平均运行长度 ARL_0 接近于 200，比两种方法失控时的平均运行长度 ARL_1，ARL_1 越小说明报警越早，监控效果越好。

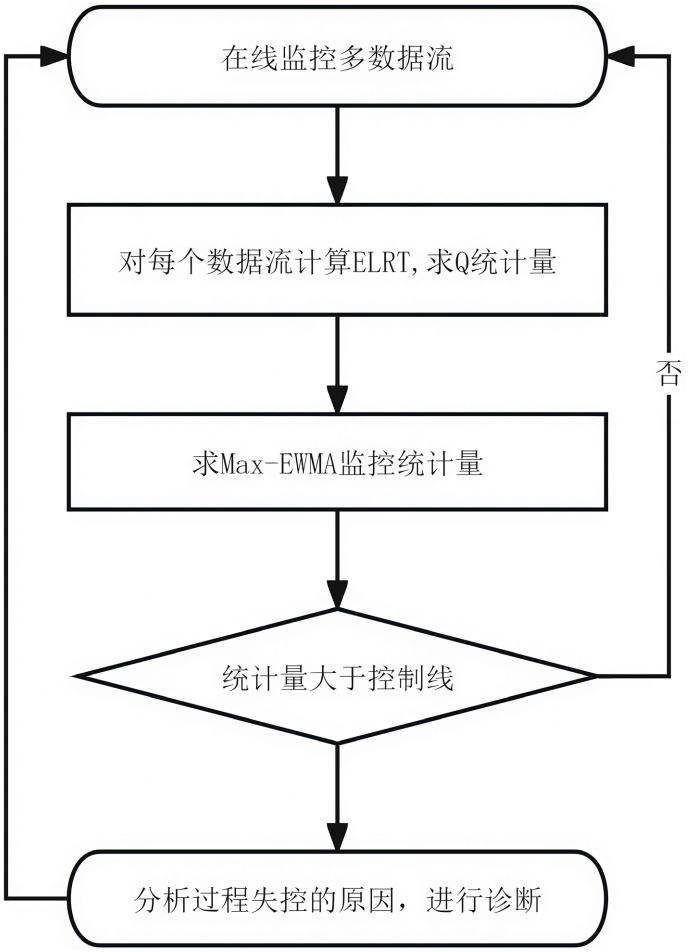

Figure 1. Flowchart for online monitoring of multiple data stream
图 1. 在线监控多数据流的流程图

为简单起见，统计模拟时假设 $N=16$，即有 16 条数据流需要监控。过程可控时，设有 5 个一元数据流服从形状参数为 5，尺度参数为 1 的威布尔分布；有 5 个一元数据流服从参数 0 和 1 的对数正态分布；有 5 个一元数据流服从自由度为 3 的 t 分布；有 1 个 5 维数据流，其边际分布服从参数为 1 的指数分布，且相互独立。样本容量 m_n 取为 30，将每条数据流都进行标准化后通过二分法模拟控制线。变点时刻 τ_n 取为 0，数据流的失控比例记为 π，漂移量记为 δ (可控数据标准化后加上 δ 即为失控数据)。在失控比例 $\pi=0.25,0.5,0.75$ 等情况下进行比较，Max-EWMA 方法都能够较早地报警，对中小漂移较为有效。

表 1 给出了失控比例为 0.25 的情况下，不同控制图的 ARL。$\delta=0$ 时的 ARL 即是 ARL_0，$\delta>0$ 时的 ARL 是 ARL_1。从表 1 可以看出，Max-EWMA 方法的表现受到光滑参数 λ 的影响，较大的光滑参数对于监控较大的漂移更有效，较小的光滑参数对于监控较小的漂移更有效。例如，$\delta=0.3$ 时，$\lambda=0.05$ 的 Max-EWMA 的 $ARL_1=24.001$，而 $\lambda=0.2$ 的 Max-EWMA 的 $ARL_1=32.259$；$\delta=0.8$ 时，$\lambda=0.05$ 的 Max-EWMA 的 $ARL_1=5.314$，而 $\lambda=0.2$ 的 Max-EWMA 的 $ARL_1=2.942$。从表 1 还可以看出，对中小漂移 Max-EWMA 方法的 ARL_1 比 Max-Q 的 ARL_1 更小，从而 Max-EWMA 方法具有更好的性能。

表 2 和表 3 分别给出了失控比例为 0.5 和 0.75 的情况下，不同控制图的 ARL。结合表 1，不难发现随着失控比例的增加，Max-EWMA 和 Max-Q 方法的报警时间越来越早。当 $\pi=0.75$ 时，ARL_1 已经很小，特别在 $\delta>0.6$ 时，ARL_1 越来越接近于 1。

Table 1. Comparison of ARL values of different control charts when $\pi = 0.25$
表 1. $\pi = 0.25$ 时不同控制图 ARL 的对比

δ	Max-EWMA		Max-Q
	$\lambda = 0.05$	$\lambda = 0.2$	
	h = 1.111718	h = 2.111279	h = 7.006011
0	200	200	200
0.3	24.001	32.259	181.764
0.4	13.789	9.495	137.759
0.5	9.923	5.778	77.571
0.6	7.859	4.276	35.84
0.7	6.418	3.455	16.191
0.8	5.314	2.942	7.314
0.9	4.268	2.512	4.024
1	3.368	2.132	2.338
2	1	1	1

Table 2. Comparison of ARL values of different control charts when $\pi = 0.5$
表 2. $\pi = 0.5$ 时不同控制图 ARL 的对比

δ	Max-EWMA		Max-Q
	$\lambda = 0.05$	$\lambda = 0.2$	
	h = 1.111718	h = 2.111279	h = 7.006011
0	200	200	200
0.3	15.484	10.403	98.773
0.4	9.741	5.413	33.565
0.5	6.984	3.751	11.284
0.6	5.032	2.864	4.619
0.7	3.507	2.251	2.407
0.8	2.391	1.803	1.655
0.9	1.758	1.496	1.229
1	1.328	1.251	1
2	1	1	1

Table 3. Comparison of ARL values of different control charts when $\pi = 0.75$
表 3. $\pi = 0.75$ 时不同控制图 ARL 的对比

δ	Max-EWMA		Max-Q
	$\lambda = 0.05$	$\lambda = 0.2$	
	h = 1.111718	h = 2.111279	h = 7.006011
0	200	200	200
0.3	7.486	4.539	24.824

续表

0.4	3.827	2.507	3.891
0.5	1.811	1.508	1.397
0.6	1.12	1.11	1.037
0.7	1.005	1.005	1
0.8	1	1	1
0.9	1	1	1
1	1	1	1
2	1	1	1

为了测试当数据流存在弱相关或强相关时 Max-EWMA 方法的性能衰减情况，假设前 15 条数据流两两之间的相关系数为 ρ，第 16 条多维数据流与其它数据流相互独立，在 $\rho=0.2$、$\rho=0.5$ 和 $\rho=0.8$ 等情况下进行统计模拟，比较 Max-EWMA 和 Max-Q 方法的性能。仅给出失控比例 $\pi=0.25$ 的情况，如表 4 所示。在失控状态下，Max-EWMA 方法展现出对相关性良好的鲁棒性。

Table 4. Comparison of ARL values of different control charts when the data stream is correlated and $\pi=0.25$
表 4. 数据流具有相关性且 $\pi=0.25$ 时不同控制图 ARL 的对比

ρ	δ	Max-EWMA		Max-Q
		$\lambda=0.05$	$\lambda=0.2$	
		h = 1.122000	h = 2.111265	h = 7.005400
0.2	0	200.72	201.18	200.28
	0.3	12.341	8.968	137.243
	0.5	4.229	2.515	6.769
	0.7	1.438	1.226	1.666
	1	1.144	1.118	1.515
		h = 1.124000	h = 2.111275	h = 7.004200
0.5	0	200.07	200.4	200.302
	0.3	12.601	9.25	139.076
	0.5	4.272	2.603	8.141
	0.7	1.505	1.201	1.513
	1	1.179	1.108	1.116
		h = 1.126000	h = 2.114800	h = 6.964600
0.8	0	200.4	200.08	199.544
	0.3	12.949	10.365	139.617
	0.5	4.459	2.585	8
	0.7	1.501	1.271	1.787
	1	1.245	1.122	1.159

理论上，当第 n 条数据流可控时，统计量 Q_{nt} 是渐近正态的，因此实际应用时需要样本容量 m_n 充分

大。经统计模拟发现，在 m_n 分别等于 20、30 和 60 的情况下，随着样本容量的增加， Q_{nt} 越来越接近于正态分布，其均值越来越接近于零。考虑到经验似然比检验的计算量及多数据流在线监控的实时性，建议实际应用中取 m_n 等于 30 或 50。

5. 结论

本文研究了基于经验似然比检验的多数据流在线监控方法，给出了监控均值或均值向量是否发生漂移的一般框架。首先提出了复杂结构的多数据流监控问题，数据流既有一元的，又有多元的；既有离散型的，又有连续型的；既有对称分布的，又有偏态分布的。利用经验似然比检验，提高了监控统计量对离散分布、偏态分布、长尾分布或未知分布的稳健性。通过 Q 统计量与 Max-EWMA 方法的结合解决了多数据流同时在线监控的问题，并对中小漂移有较好的表现。以不同的失控比例，通过统计模拟分析了所提出方法在不同漂移量下的失控时的平均运行长度。实验结果表明所提出的在线监控方法具有较好的性能。本文的研究假设数据流之间是相互独立的，当控制图的方法发出失控警报时可以利用 Q 统计量的 EWMA 值在哪个数据流达到最大进行诊断，判断到底哪条或哪些数据流失控了。在之后的研究中可以考虑数据流之间具有一定的相关性时，数据流的个数是可变的情况下讨论在线监控及诊断问题。

基金项目

国家自然科学基金面上项目(12271271)。

参考文献

[1] 朱建平, 冯冲, 梁振杰. 交叉学科促进统计学的发展[J]. 统计研究, 2023, 40(1): 134-144.

[2] Mei, Y. (2010) Efficient Scalable Schemes for Monitoring a Large Number of Data Streams. *Biometrika*, **97**, 419-433. https://doi.org/10.1093/biomet/asq010

[3] Du, L. and Zou, C. (2018) Online Control of False Discovery Rates for Multiple Datastreams. *Journal of Statistical Planning and Inference*, **194**, 1-14. https://doi.org/10.1016/j.jspi.2017.10.006

[4] Ren, H., Zou, C., Chen, N. and Li, R. (2020) Large-Scale Datastreams Surveillance via Pattern-Oriented-Sampling. *Journal of the American Statistical Association*, **117**, 794-808. https://doi.org/10.1080/01621459.2020.1819295

[5] Dassanayake, S. and French, J.P. (2023) Detecting Disease Outbreak Regions Using Multiple Data Streams. *Statistics in Biosciences*, **16**, 142-164. https://doi.org/10.1007/s12561-023-09387-5

[6] Zhou, Q., Zou, C., Wang, Z. and Jiang, W. (2012) Likelihood-Based EWMA Charts for Monitoring Poisson Count Data with Time-Varying Sample Sizes. *Journal of the American Statistical Association*, **107**, 1049-1062. https://doi.org/10.1080/01621459.2012.682811

[7] Li, J., Tsung, F. and Zou, C. (2014) Multivariate Binomial/Multinomial Control Chart. *IIE Transactions*, **46**, 526-542. https://doi.org/10.1080/0740817x.2013.849830

[8] 范倩. 基于似然比检验的对数正态分布控制图[D]: [硕士学位论文]. 沈阳: 辽宁大学, 2018.

[9] 郭宝才, 李敏, 项朝辉, 等. 参数未知下基于定数截尾样本监控指数分布的控制图设计[J]. 高校应用数学学报 A 辑, 2018, 33(1): 1-12.

[10] 曹程明. 基于威布尔分布的高质量过程统计控制研究[D]: [硕士学位论文]. 南京: 南京理工大学, 2019.

[11] Sanusi, R.A., Teh, S.Y. and Khoo, M.B.C. (2020) Simultaneous Monitoring of Magnitude and Time-Between-Events Data with a Max-EWMA Control Chart. *Computers & Industrial Engineering*, **142**, Article ID: 106378. https://doi.org/10.1016/j.cie.2020.106378

[12] 刘英杰. 基于多元离散型数据的 CUSUM 控制图研究[D]: [硕士学位论文]. 天津: 天津职业技术师范大学, 2022.

[13] 廉惠然, 齐德全. 统计过程控制在网络销售中的应用[J]. 应用数学进展, 2023, 12(2): 609-614.

[14] 王兆军, 邹长亮, 李忠华. 统计质量控制图理论与方法[M]. 北京: 科学出版社, 2013.