Statistics and Application 统计学与应用, 2024, 13(2), 307-314
Published Online April 2024 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2024.132031

Hans汉斯

基于错误发现率的高维数据流在线监控方法

梁　楠，齐德全*

长春理工大学数学与统计学院，吉林 长春

收稿日期：2024年3月11日；录用日期：2024年4月1日；发布日期：2024年4月9日

摘　要

关于多数据流的监控，大多假设数据流之间是独立的。从统计过程控制的角度，给出了在线监控高维数据流的一般框架。鉴于数据的分布可能存在多样性，本文采用对称数据聚合方法建立了稳健的监控统计量，利用统计量的渐进对称性选取数据驱动的阈值，基于错误发现率对相关的非正态数据流进行在线监控。以AR (1)模型刻画数据流间的相关性，通过蒙特卡洛模拟，研究了所提出方法的错误发现率和功效水平。数值模拟结果表明所提出的方法具有较理想的性能。

关键词

错误发现率，对称数据聚合，高维数据流，统计过程控制

Online Monitoring Method of High-Dimensional Data Streams Based on False Discovery Rate

Nan Liang, Dequan Qi*

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

Received: Mar. 11th, 2024; accepted: Apr. 1st, 2024; published: Apr. 9th, 2024

Abstract

Regarding the monitoring of multiple data streams, it is mostly assumed that the data streams are independent. A general framework for online monitoring of high-dimensional data streams is provided from the perspective of statistical process control. Given the potential diversity in data distribution, this paper adopts a symmetric data aggregation method to establish a robust moni-

*通讯作者。

文章引用: 梁楠, 齐德全. 基于错误发现率的高维数据流在线监控方法[J]. 统计学与应用, 2024, 13(2): 307-314.
DOI: 10.12677/sa.2024.132031

toring statistic. The asymptotic symmetry of the statistic is used to select data-driven thresholds, and the relevant non-normal data streams are monitored online based on the false discovery rate. The AR (1) model was used to characterize the correlation between data streams, and the false discovery rate and power level of the proposed method were studied through Monte Carlo. The numerical simulation results indicate that the proposed method has ideal performance.

Keywords

False Discovery Rate, Symmetric Data Aggregation, High-Dimensional Data Streams, Statistical Process Control

1. 引言

随着数据科学时代的到来，作为保证产品与服务符合规定的要求的一种质量管理技术，统计过程控制(Statistical Process Control, SPC)的理论研究越来越丰富，已由监控一元数据到多元数据[1] [2]，由点数据到线数据[3]及面数据[4]，由多数据流[5]发展到高维数据流[6]。Spiegel Halter 等人[6]考虑英国医疗保健系统中超额死亡率的 200,000 个指标，建立了 200,000 个独立的累积和(Cumulative sum, CUSUM)控制图用来监控这些指标。

控制图的方法通常被用来监控高维数据流均值向量[7]或协方差矩阵的漂移[8]。当监控的数据流有成千上万条时，很可能在每个数据收集的时间段内都会发出过程失控的报警信号。于是传统的控制图评判指标，如错误报警率(False Alarm Rate, FAR)和平均运行长度(Average Run Length, ARL)等都失去了效用。因此 Qi 等人[9]基于错误发现率(False Discovery Rate, FDR)研究了高维数据流的数据质量的在线监控方法。假设相互独立的数据流的个数充分大，每个数据流是一元的，观察其准确性、一致性和完备性等多个数据质量(多元离散型向量，其边际分布是 0~1 分布，具有一定的相关性)，监控数据质量的均值向量是否发生变化。与文献[5] [6] [7]不同，Qi 等人[9]监控均值向量时假设数据流的个数是随着监控时间变化的，监控过程是可出可进的，发出过程失控警报的数据流经诊断之后可以重新回到监控系统，也可以在监控过程中加入新的未监控过的数据流。但 Qi 等人[9]所提出的控制 FDR 方法没能证明出监控统计量的分布，在计算监控统计量的 p 值时是通过统计模拟实现的(类似于 Shen 等人[10]的基于模拟的方法)。这样带来沉重的计算负担，影响在线监控的实时性。

FDR 始于 Benjamini 和 Hochberg [11]的开创性工作。为了提高多重假设检验的功效，Benjamini 和 Hochberg 提出了 FDR，并在独立的情况下给出了控制 FDR 的 BH 方法。之后，学者们集中于相关情况[12] [13]、稳健方法[14] [15]和各领域应用[16] [17] [18]等方面的研究。Du 等人[19]把 SPC 与变量选择和 FDR 相结合，在具有一定相关性的情况下，基于对称数据聚合(Symmetric Data Aggregation, SDA) [20]研究了适用于非正态高维数据流的稳健监控方法。对于每一维度，通过求均值构造对称统计量，进一步求得 SDA 监控统计量。当 SDA 的某个分量大于数据驱动的阈值时，监控过程发出该分量所对应的数据流失控的警报。Du 等人[19]证明了所提出的 SDA 方法能把 FDR 控制在指定范围之内。相较于 Qi 等人[9]的方法，Du 等人[19]的 SDA 方法只需要统计量满足对称性就能控制住 FDR，不需要经统计模拟求统计量的 p 值再由 BH 方法控制 FDR，计算量会显著下降，因此监控的实时性会更好。但 Du 等人[19]没有讨论数据流

可进可出的在线监控情况。

综上，本文考虑以下复杂结构的高维数据流可进可出的在线监控问题，仅给出监控均值向量是否发生漂移的一般框架。数据流的个数充分大，部分数据流之间具有一定的相关性，单个数据流可以是一元的也可以是多元的，可以是连续的随机变量也可以是离散的随机变量。结合多元指数加权滑动平均(Multivariate Exponentially Weighted Moving Average, MEWMA)统计量与 SDA 方法采用数据驱动的阈值控制 FDR，实现对具有一定相关性的高维数据流的在线监控。通过蒙特卡洛模拟，所提出的方法能够把 FDR 控制到目标水平以下，同时具有较高的功效。

2. 变点模型

随着大数据的不断涌现，假设在生产或服务等过程中，在时刻 t 需要监控 N_t 条数据流 $\boldsymbol{X}_1,\cdots,\boldsymbol{X}_{N_t}$。这里 N_t 是取值充分大的正整数，第 n 条数据流 $\boldsymbol{X}_n$ 根据问题背景的不同取值为一元随机变量或多元随机向量，其均值或均值向量记为 $\boldsymbol{\mu}_n$，其方差或协方差矩阵记为 $\boldsymbol{\Sigma}^{(n)}$。在 N_t 条数据流中，根据实际情况的不同，可能既有一元的数据流又有多元的数据流，也可能仅有其中的一种。数据流之间的相关性通过协方差矩阵 $\boldsymbol{\Sigma}$ 来刻画，并假设这个相关性在整个监控过程中不变。假设存在一个未知的时刻 τ_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：

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

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

若 $t+1$ 时刻有新的数据流被加入到监控系统，则数据流的个数 N_{t+1} 相应地增加。若当前时刻 t，根据下面的在线监控方法得出某些数据流发出失控的警报，则下一时刻 N_{t+1} 相应地减少。

3. 在线监控方法

在时刻 t，对第 n 条数据流抽取 m_n 个样本 $\left(X_{n,t,1},\cdots,X_{n,t,m_n}\right)$。鉴于数据流变量维数和类型的复杂结构，结合 MEWMA 统计量[1]与 SDA 方法[19]进行在线监控。首先将 m_n 个样本分成两个不相交的容量分别为 m_{n1} 和 $m_{n2}=m_n-m_{n1}$ 的子集 D_{n1} 和 D_{n2}，计算渐进对称分布的统计量 $T_{n,t}^1$ 和 $T_{n,t}^2$。若 $\boldsymbol{X}_n$ 是一元随机变量，则

$$T_{n,t}^b=\frac{\sum_{i\in D_{nb}}X_{n,t,i}}{\sqrt{m_{nb}}},b=1,2.$$

若 $\boldsymbol{X}_n$ 是多元随机向量，则先求 MEWMA 统计量

$$Z_{m_{nb}}=(1-\lambda)Z_{m_{nb}-1}+\lambda\left(X_{n,t,m_{nb}}-\mu_n^0\right),b=1,2,$$

其中 $Z_0=0$，$\lambda\in(0,1)$ 是光滑参数。可以证明 $Z_{m_{nb}}$ 的均值向量为零向量，协方差矩阵为 $\frac{\lambda\left(1-(1-\lambda)^{2m_{nb}}\right)}{2-\lambda}\Sigma^{(n)}$，从而

$$T_{n,t}^b=\frac{2-\lambda}{\lambda\left(1-(1-\lambda)^{2m_{nb}}\right)}Z'_{m_{nb}}\left(\Sigma^{(n)}\right)^{-1}Z_{m_{nb}},b=1,2.$$

当 m_{n1} 和 m_{n2} 较大时，$T_{n,t}^1$ 和 $T_{n,t}^2$ 可以近似计算为 $T_{n,t}^b=\frac{2-\lambda}{\lambda}Z'_{m_{nb}}\left(\Sigma^{(n)}\right)^{-1}Z_{m_{nb}},b=1,2$。由中心极限定理得，在数据流可控时，上面所求的 $T_{n,t}^1$ 和 $T_{n,t}^2$ 渐进服从正态分布，因此是对称的统计量。

然后构造 SDA 监控统计量 $W_{n,t}=T_{n,t}^1\cdot T_{n,t}^2$。当 $W_{n,t}\geq L_t$ 时，发出第 n 条数据流过程失控的警报，其中数据驱动的阈值 $L_t=\inf\left\{r_t>0:\frac{\#\{n:W_{n,t}\leq -r_t\}}{\#\{n:W_{n,t}\geq r_t\}\vee 1}\leq\alpha\right\}$，$\alpha$ 是事先给定的希望 FDR 所满足的水平。

监控过程的流程图如图 1 所示。

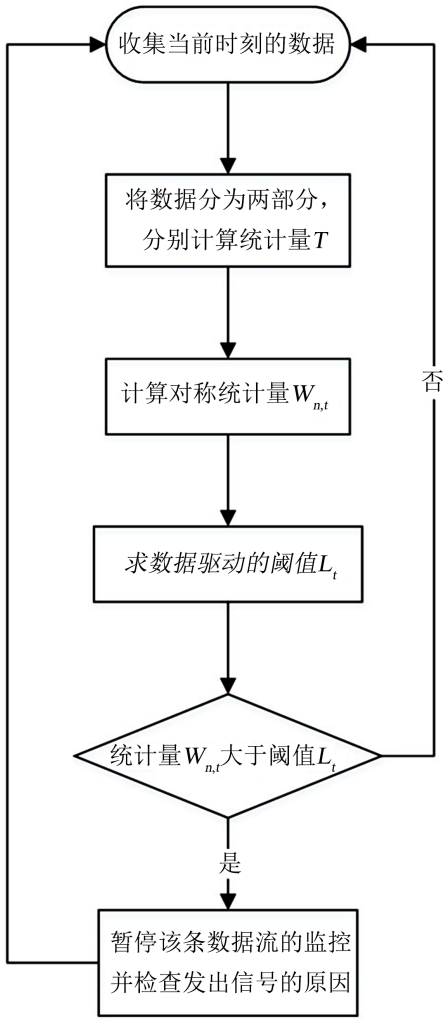

Figure 1. Flowchart for online monitoring of high-dimensional data stream
图 1. 在线监控高维数据流的流程图

4. 统计模拟与示例

通过蒙特卡洛模拟说明所提出方法(简记为 OSDA)监控非正态高维数据流的有效性，通过意大利 GIMEMA 临床试验的探针集数据示例 OSDA 方法的应用。

为简单起见，统计模拟时假设每个数据流都服从一元 t 分布，数据流之间的协方差矩阵满足 $\Sigma=\left(\rho^{|j-k|}\right)_{N_t\times N_t}$，$j,k=1,2,\cdots,N_t$，$\rho>0$ 为常数。当数据流可控时 $\mu_n^0=0$，数据流失控时 $\mu_n^1=A$，这里 A 表示发生的漂移量，失控时数据流变为非中心的 t 分布，数据流的失控比例记为 π。进一步假设 $N_1=1000$，$\alpha=0.2$，$m_n=90$。参考文献[19]和[20]，取 $m_{n1}=2m_n/3=60$。

因为每个数据流都是一元的，监控其均值是否发生漂移，所以仿照文献[19]将 t 检验和 BH 方法相结合作为对比方案(简记为 BH)。进行 500 次重复模拟实验，对比两种方法在相同条件下的 FDR 和功效 Power。在 FDR 不超过预设的上限 α 的情况下，方法的 Power 值越大，说明该方法越有效。在 t 分布的自由度 $df=3,4,5,6,7$，$\rho=0.2,0.4,0.5,0.6,0.8$，漂移大小 $A=0.2,0.3,0.4,0.5$，失控比例 $\pi=0.3,0.4,0.5,0.6$ 等情况

下进行比较，OSDA 方法都够把 FDR 控制在目标水平以下，且具有较高的功效。

为了说明在相同的自由度和失控比例的情况下，不同的漂移量对这两种方法性能的影响。图 2 给出了自由度 $df=3$，失控比例 $\pi=0.5$，漂移量 $A=0.3$ 和 $A=0.4$ 的情况下，两种方法的比较结果。由图 2(a)和图 2(b)的对比可以看出，两种方法都可以将 FDR 控制在目标水平下，OSDA 方法得到的 FDR 值可以更好的靠近 FDR 目标水平，BH 方法得出的是更为保守的值。由图 2(c)和图 2(d)的对比可以看出，OSDA 方法在不同漂移下的累积功效均高于 BH 方法，而且在较小漂移情况下，OSDA 方法的优势更为明显。

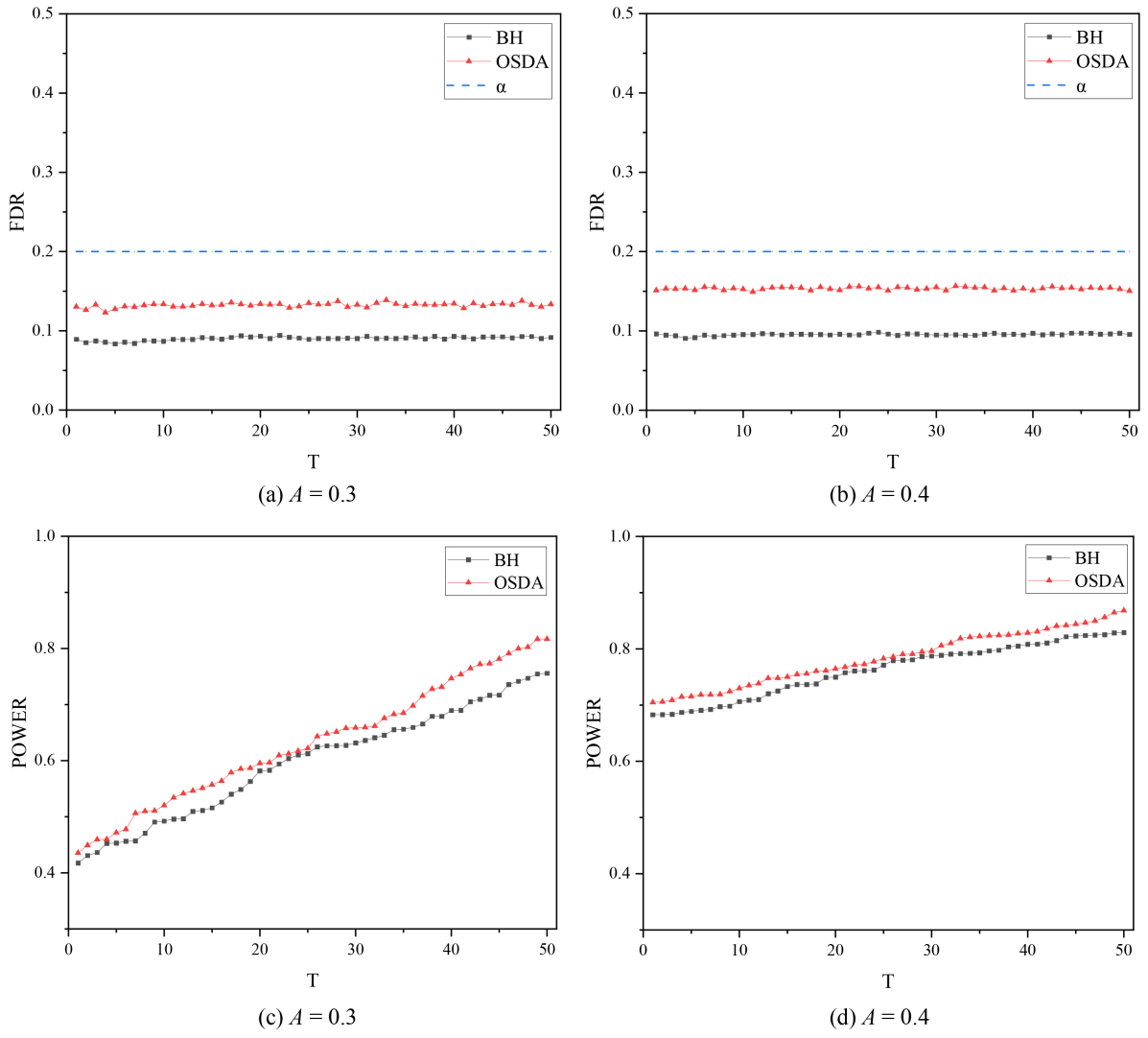

Figure 2. FDR and AP of online monitoring of multivariate *t*-distribution data with different signal amplitude
图 2. 在线监控不同漂移的多元 *t* 分布数据的错误发现率和累积功效

为了说明在相同的自由度和漂移情况下，不同的失控比例对这两种方法性能的影响。图 3 给出了自由度 $df=3$，漂移量 $A=0.3$，失控比例 $\pi=0.4$ 和 $\pi=0.6$ 的情况下，两种方法的比较结果。由图 3(a)和图 3(b)的对比可以看出，两种方法都可以将 FDR 控制在目标水平之下，OSDA 方法得到的 FDR 值更好的靠近 FDR 目标水平，BH 方法得出的是更为保守的值。由图 3(c)和图 3(d)的对比可以看出，OSDA 方法在不同漂移下的累积功效均高于 BH 方法，而较大失控比例情况下，OSDA 方法的优势更为明显。

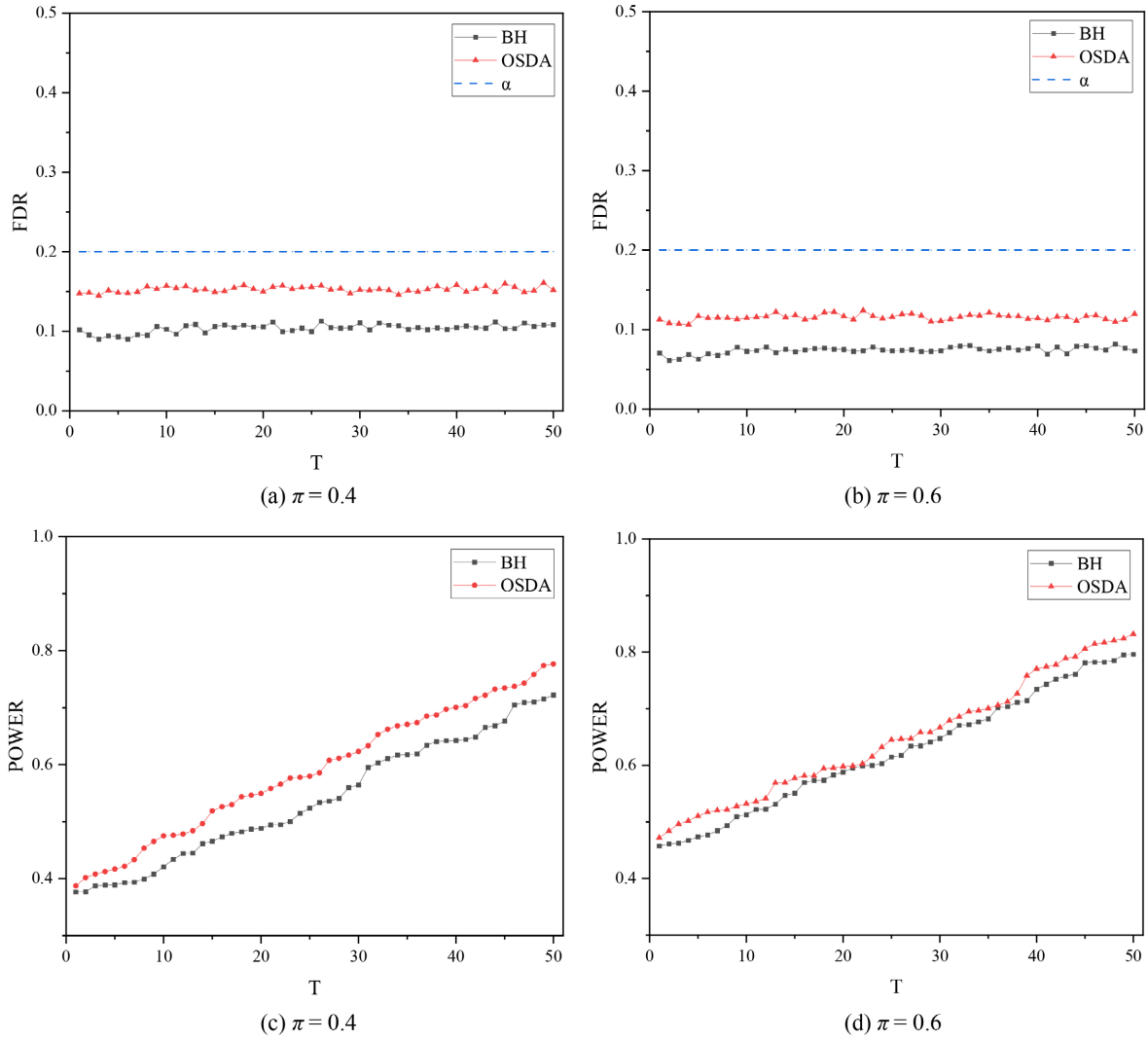

Figure 3. FDR and AP of online monitoring of multivariate t-distribution data with different out of control ratio
图 3. 在线监控不同失控比例的多元 t 分布数据的错误发现率和累积功效

为了评估两种方法的性能，还比较了$(A,\ \pi)$不同组合下 FDR 的平均绝对误差(Average Absolute Error, AAE)和累积功效。表 1 仅给出了自由度 $df=3$，$\rho=0.8$ 时的比较结果。表中的 AAE 是由时刻 $t=1$ 至 $t=100$，$FDR_t-\alpha$ 的绝对值的平均数，表中的 Power 是在时刻 $t=100$ 处的累积功效。

Table 1. Compare OSDA and BH methods under different combinations of (A, π)
表 1. 在(A, π)的不同组合下对比 OADA 和 BH 方法

A		$\pi=0.4$		$\pi=0.5$		$\pi=0.6$	
		AAE	Power	AAE	Power	AAE	Power
0.3	OSDA	0.048	0.772	0.068	0.819	0.081	0.852
	BH	0.092	0.721	0.110	0.757	0.127	0.793
0.4	OSDA	0.028	0.807	0.049	0.868	0.051	0.925
	BH	0.092	0.759	0.104	0.834	0.122	0.861

从表 1 中可以看出，与 BH 方法相比，OSDA 方法具有更小的 AAE，更大的 Power 值，因此监控效果更好。

Du 等人[19]的实例分析使用了意大利 GIMEMA 临床试验的 1263 个基因探针集。本文在 1263 个基因探针集中随机抽取 1000 个进行在线监控作为示例。取 $\alpha = 0.2$，在每个数据流上抽取 79 个样本(b 细胞分化患者)，分割为两个大小分别为 37 和 42 的子集，计算 SDA 监控统计量 $W_{n,t}$，根据阈值 L_t 判断某数据流是否失控。监控效果如图 4 所示。

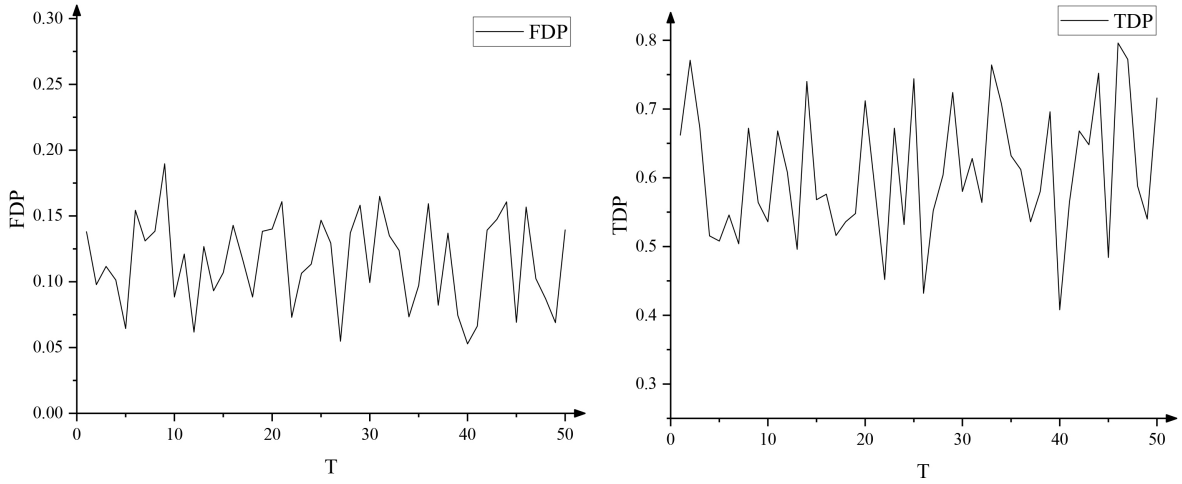

Figure 4. The FDP and TDP level of the OSDA method once test
图 4. OSDA 方法单次检验的错误发现比例和正确发现比例

因为只运用 OSDA 方法做一次在线监控，所以图 4 中计算的是错误发现比例(False Discovery Proportion, FDP)和正确发现比例(True Discovery Proportion, TDP)。由图 4 可以看出，FDP 值都被控制在预先设定的水平之下，而且功效较好。可见，所提出的 OSDA 方法具有一定的适用性。

5. 结论

本文研究了基于错误发现率的高维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高维数据流监控问题，数据流变量的维数和类型是多样的，监控过程中数据流是可出可进的。利用 SDA 方法控制错误发现率，解决了前人工作中基于模拟方法控制错误发现率的计算量问题，使得在线监控更具有时效性。通过 MEWMA 统计量与 SDA 方法的结合解决了复杂结构数据流的监控问题。通过统计模拟分析了所提出方法的错误发现率、功效和平均绝对误差等指标。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具有一定的相关性，在整个监控过程中刻画相关性的协方差阵是不变的，在之后的研究中可以考虑监控协方差阵的漂移或研究其影响。

基金项目

吉林省教育厅项目(JJKH20210809KJ)、国家自然科学基金面上项目(12271271)。

参考文献

[1] Bersimis, S., Psarakis, S. and Panaretos, J. (2007) Multivariate Statistical Process Control Charts: An Overview. *Quality and Reliability Engineering International*, **23**, 517-543. https://doi.org/10.1002/qre.829

[2] Woodall, W.H. and Montgomery, D.C. (2014) Some Current Directions in the Theory and Application of Statistical Process Monitoring. *Journal of Quality Technology*, **46**, 78-94. https://doi.org/10.1080/00224065.2014.11917955

[3] Noorossana, R., Saghaei, A. and Amiri, A. (2011) Statistical Analysis of Profile Monitoring. John Wiley & Sons, Inc., Hoboken. https://doi.org/10.1002/9781118071984

[4] Wang, A., Wang, K. and Tsung, F. (2014) Statistical Surface Monitoring by Spatial-Structure Modeling. *Journal of Quality Technology*, **46**, 359-376. https://doi.org/10.1080/00224065.2014.11917977

[5] 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

[6] Spiegelhalter, D., Sherlaw-Johnson, C., Bardsley, M., Blunt, I., Wood, C. and Grigg, O. (2012) Statistical Methods for Healthcare Regulation: Rating, Screening and Surveillance (with Discussions). *Journal of the Royal Statistical Society Series A*, **175**, 1-47. https://doi.org/10.1111/j.1467-985X.2011.01010.x

[7] Zou, C., Wang, Z., Zi, X., *et al*. (2015) An Efficient Online Monitoring Method for High-Dimensional Data Streams. *Technometrics*, **57**, 374-387. https://doi.org/10.1080/00401706.2014.940089

[8] Kim, J., Abdella, G.M., Kim, S., *et al*. (2019) Control Charts for Variability Monitoring in High-Dimensional Processes. *Computers & Industrial Engineering*, **130**, 309-316. https://doi.org/10.1016/j.cie.2019.02.012

[9] Qi, D., Li, Z. and Wang, Z. (2016) On-Line Monitoring Data Quality of High-Dimensional Data Streams. *Journal of Statistical Computation and Simulation*, **86**, 2204-2216. https://doi.org/10.1080/00949655.2015.1106542

[10] Shen, X., Zou, C., Jiang, W. and Tsung, F. (2013) Monitoring Poisson Count Data with Probability Control Limits When Sample Sizes Are Time Varying. *Naval Research Logistics*, **60**, 625-636. https://doi.org/10.1002/nav.21557

[11] Benjamini, Y. and Hochberg, Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society, Series B*, **57**, 289-300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

[12] Finner, H., Dickhaus, T. and Roters, M. (2007) Dependency and False Discovery Rate: Asymptotics. *The Annals of Statistics*, **35**, 1432-1455. https://doi.org/10.1214/009053607000000046

[13] Fan, J. and Han, X. (2017) Estimation of the False Discovery Proportion with Unknown Dependence. *Journal of the Royal Statistical Society Series B*: *Statistical Methodology*, **79**, 1143-1164. https://doi.org/10.1111/rssb.12204

[14] He, Y., Zhang, X., Wang, P., *et al*. (2017) High Dimensional Gaussian Copula Graphical Model with FDR Control. *Computational Statistics & Data Analysis*, **113**, 457-474. https://doi.org/10.1016/j.csda.2016.06.012

[15] Yuan, P., Kong, Y. and Li, G. (2023) FDR Control and Power Analysis for High-Dimensional Logistic Regression via StabKoff. *Statistical Papers*. https://doi.org/10.1007/s00362-023-01501-5

[16] Barras, L., Scaillet, O. and Wermers, R. (2010) False Discoveries in Mutual Fund Performance: Measuring Luck in Estimated Alphas. *The Journal of Finance*, **65**, 179-216. https://doi.org/10.1111/j.1540-6261.2009.01527.x

[17] Schwartzman, A., Dougherty, R.F. and Taylor, J.E. (2008) False Discovery Rate Analysis of Brain Diffusion Direction Maps. *The Annals of Applied Statistics*, **2**, 153-175. https://doi.org/10.1214/07-AOAS133

[18] Sun, W., Reich, B.J., Tony, C.T., *et al*. (2015) False Discovery Control in Large-Scale Spatial Multiple Testing. *Journal of the Royal Statistical Society Series B*: *Statistical Methodology*, **77**, 59-83. https://doi.org/10.1111/rssb.12064

[19] Du, L., Guo, X., Sun, W., *et al*. (2023) False Discovery Rate Control under General Dependence by Symmetrized Data Aggregation. *Journal of the American Statistical Association*, **118**, 607-621. https://doi.org/10.1080/01621459.2021.1945459

[20] Wasserman, L. and Roeder, K. (2009) High Dimensional Variable Selection. *Annals of Statistics*, **37**, 2178-2201. https://doi.org/10.1214/08-AOS646