一种稳健多变量自相关控制图设计方法及其在空气质量监控中的应用

王志坚[1,2]，张钰哲[1]，杜玮[1]
1. 广东财经大学 统计与数据科学学院，广东 广州 510320
2. 新疆政法学院 经济管理学院，新疆 图木舒克 843900
DOI:10.61369/ASDS.2026100008

摘　　要： 传统 VAR 控制图面对含异常值的多变量时间序列时，往往容易被个别异常点带偏，且往往无法准确指出所有发生异常的变量。监控结果一旦与真实情况错位，决策者就很难及时锁定并调整异常变量。针对这一短板，本文把稳健估计量与加权方法 Zt 结合起来，构造了一种稳健 VAR 控制图，并通过不同污染率、不同维度下的模拟实验检验其表现。结果显示，与原有方法相比，新控制图在异常识别的准确率和灵敏度上都有改善，异常值比例较高时优势更明显。这项工作为多变量时间序列监控增加了一种可选工具，也为数据驱动决策中的稳健性分析提供了参考。

关 键 词： 向量自回归控制图；高维数据处理；异常值检测；过程控制

A Design Method for Robust Multivariate Autocorrelation Control Charts and Its Application in Air Quality Monitoring

Wang Zhijian[1,2], Zhang Yuzhe[1], Du Wei[1]
1.School of Statistics and Data Science , Guangdong University of Finance & Economics, Guangzhou, Guangdong 510320
2.School of Economics and Management, Xinjiang University of Political Science and Law, Tumxuk, Xinjiang 843900

Abstract: Traditional VAR control charts tend to be pulled off course by individual outliers when monitoring multivariate time series containing anomalies, and they often fail to clarify which variables actually went wrong. Once monitoring results deviate from the real situation, decision-makers struggle to pinpoint and adjust anomalous variables in time. To address this weakness, this paper combines robust estimators with the weighted Ztmethod and develops a robust VAR control chart. Simulation experiments under different contamination rates and dimensions are used to test its performance. The results show that, compared with the original method, the new chart improves both accuracy and sensitivity in outlier detection, and the advantage is clearer when the proportion of outliers is high. This work provides an additional tool for multivariate time series monitoring and a reference for robustness analysis in data-driven decision-making.

Keywords: vector autoregression control charts; high-dimensional data processing; outlier detection; process control

引言

统计过程控制（Statistical Process Control, SPC）一直是质量管理的基础工具，在自动化程度高、、制造环节复杂的生产现场尤为常见。多变量统计过程控制（Multivariate Statistical Process Control, MSPC）能同时盯住多个质量特性，因此在复杂生产过程中很难被替代。不过，工业数据的复杂度和维度都在上升，传统多变量控制图用起来越来越吃力。

问题出在两个地方。其一，许多工业过程的数据里，当前观测值和过去观测值并不独立，也就是存在自相关。传统控制图偏偏假设观测独立，这个假设在实际数据里根本不成立，控制图就容易被过程偏差带偏，虚警和漏报都会冒出来。其二，工业数据中还常混有异常值，它们可能来自测量误差、设备故障或突发事故。这些异常点一旦掺进来，均值和协方差的估计就会被拉动，报警限随之失真，整个质量监控系统的可靠性也跟着打折扣。

基金项目：本文获得广东省基础与应用基础研究基金项目“多变量自相关过程稳健控制设计及评价方法研究”（项目编号：2026A1515011461）资助。
作者简介：
王志坚（1982-），男，江西余干人，博士，副教授，研究方向：统计过程控制、时间序列分析；
张钰哲（2000-），男，湖南长沙人，硕士，研究方向：时间序列分析；
杜玮（2005—），女，河南驻马店人，硕士研究生在读，研究方向：时间序列分析。

Copyright © This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

为了克服这些问题，近年来，稳健统计方法在多变量控制图中的应用得到了广泛关注，取得了较好的研究成果。理论层，Centofanti 等构建的功能回归控制图突破传统有限维监控边界 [1]，Elisa 等的收缩估计方法解决高维小样本协方差矩阵估计奇异痛点 [2]，Bayati 等的非参数深度控制图、Pulido 等的谨慎学习机制，从分布假设、受控性能两个维度完成方法底层升级 [3,4]，王志坚等将稳健估计量应用于Shewhart控制图取得了良好的效果，为稳健控制图的应用提供基础支撑 [5]。方法优化层，AliH T 等拓展变异性监控能力 [6]，S V 等完善理性子组下的小样本监控体系 [7]，薛丽等首次将自相关时序特征纳入多变量控制图优化框架 [8]。应用层，Capezza 等、Will 等分别完成环境排放、生物测序场景的跨域验证 [9,10]，雷彬文等在工业生产场景完成实证落地 [11]。但现有研究均未兼顾空气质量数据极端污染值、多污染物耦合与时序自相关的叠加干扰，面向该场景的专用稳健控制图的研究仍亟待完善。

为了解决上述问题，本文提出了一种基于稳健统计和时间序列模型的稳健多变量自相关控制图设计方法。该方法结合了重加权最小协方差行列式（RMCD）和重加权最小体积椭球体（RMVE）等稳健估计量，能够有效减少异常值对均值和协方差估计的影响；同时，采用一阶向量自回归模型（VAR(1)）描述多变量数据的时间依赖性，从而提高了对过程异常变化的检测灵敏度和准确性。通过蒙特卡洛仿真实验，验证了所提出方法在高维度、强自相关及高异常值率的场景下的优越性能，尤其在早期异常检测和过程偏移诊断方面，显著提高了过程监控的效率和稳定性。

本文的工作为工业过程质量监控提供了一条新思路。借助稳健多变量自相关控制图，复杂工业数据中的异常值和自相关问题有望得到更好的处理，过程控制的可靠性和实时性也随之改善。就目前的结果看，这项工作在理论和应用层面都还有继续挖掘的空间。

一、传统向量自回归控制图的原理及其不稳健性

（一）传统向量自回归控制图的原理

要看清稳健 VAR 控制图为何有必要，得先弄明白传统 VAR 控制图是怎么运作的。

设有 p 个变量的时间序列数据，每个变量有 n 个观测值，写成向量 $X_t=\left(x_{1t},x_{2t},\cdots,x_{pt}\right)$，其中 $t=1,2,\cdots,n$，x_{it} 表示第 i 个变量在时间 t 的取值。一阶 p 维的 VAR 模型 (VAR(1)) 可以写成：

$$X_t=\mu_t+\Phi\left(X_{t-1}-\mu_t\right)+\varepsilon_t$$

其中 μ_t 为 t 时刻的均值向量，μ_t 为均值向量为零的独立正态随机变量向量，协方差矩阵 Σ，Φ 为自相关参数 $p\times p$ 矩阵。进一步假设 X_t 是平稳的。在此假设下，μ_t 随时间的变化为常数，因此可以写成：

$$X_t=\mu+\Phi\left(X_{t-1}-\mu\right)+\varepsilon_t$$

根据 A. A. Kalgonda and S. R. Kulkarni[12] 提出的 VAR 控制图构建的建模步骤可知，假定在时间点 t 观测到的 p 维随机变量 X_t 严格遵循第一阶向量自回归过程 VAR(1)，即 $X_t\sim N_p(\mu,\Gamma(0))$，其中 $\Gamma(0)$ 是交叉协方差矩阵。当过程处于控制状态时，X_t 服从均值向量为 $\mu_0=\left(\mu_{10},\mu_{20},\cdots,\mu_{p0}\right)$ 的多元正态分布，且 μ_0 是目标均值向量。

当过程失去控制时，通常可能是由于变量子集的变化，而不是所有变量同时发生变化 (Hawkins，1991[13])。因此，在这种情况下，如果至少有一种假设不是真的，则认为过程失控更为合适：

$$H_{i0}:\mu_i=\mu_{i0},i=1,2,\cdots p$$

如果上式对所有的 i 成立，则认为该过程受控，即为，

$$H_0:\bigcap H_{i0}$$

设置控制限：下控制限 (LCL) 设为 0，上控制限 (UCL) 设为 $C_{\rho(0),\alpha}$，其中 $C_{\rho(0),\alpha}$ 是基于置信水平和自相关结构确定的常数。

$$Z_t=Max_{1\le i\le p}\left[\left|Z_{it}\right|\right]$$

其中 Z_{it} 是标准化后的观测值，$\sqrt{\gamma_{ii}(0)}$ 是 $\Gamma(0)$ 对角线上的元素。

$$Z_{it}=\frac{x_{it}-\mu_{i0}}{\sqrt{\gamma_{ii}(0)}},i=1,2,\cdots,p$$

当 H_{i0} 为真时，可知 Z_{it} 服从标准正态分布。显然 Z_{it} 不是独立的，因此，当过程处于受控状态时，可知，

$$Pr_{H0}\left[\left|Z_{it}\right|\le C_{\rho(0),\alpha},i=1,2,\cdots,p\right]=1-\alpha$$

根据上式在取定置信水平 a 时，$C_{\rho(0),\alpha}$ 的估计为样本 Z_t 的第 $(1-\alpha)$ 百分位点。

到此，已经完成了 VAR 控制图的各要素计算，接下来绘制 Z_t 的散点图，判断控制状态：如果 Z_t 值在控制限 (LCL 和 UCL) 范围内，则认为过程处于控制状态；如果超出范围，则认为过程失控。一旦确定过程失控，根据 Z_t 对应的 Z_{it} 可以直接确定发生失控的变量，从而进行针对性的调整或干预。

（二）传统向量自回归控制图的不稳健性和局限性

先看参数估计这一环。传统 VAR 控制图大多借助最小二乘法或极大似然法来估计模型参数，而这两种方法都绕不开样本均值和协方差矩阵。问题在于，这两个统计量恰恰对异常值很敏感。过程数据里只要混进少量异常观测，VAR 模型的参数估计就会明显跑偏，残差序列也随之不再满足独立同分布假设，后续搭起来的控制统计量便失去了理论根基。结果就是，过程明明还稳定，控制图却可能频繁报警，虚警概率一路往上升。

异常值和自相关一旦同时出现，传统 VAR 控制图的下滑会更明显。自相关结构会充当放大器，把异常值对参数估计的干扰成倍放大，异常观测借着时间依赖关系向相邻样本扩散；与此同时，异常值又会削弱模型刻画真实自相关结构的能力，造成拟合不足或过度拟合。两股作用叠加在一起，控制限就很难如实反映过程的真实波动水平，控制图对过程状态变化的判别能力也随之

 Copyright © This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

明显下降。

为验证这一点，本文设计了一组数值模拟实验。先构建一阶向量自回归模型，生成二维多变量时间序列样本，用来刻画变量之间随时间演化的动态相关结构，其中当前观测由前一时刻状态和随机扰动共同决定。在此基础上随机生成100个观测值作为基准数据集，并从中抽取约10%的样本施加人为扰动，模拟异常值。通过分析污染数据条件下VAR控制图的监控表现，观察异常值对其稳定性和报警性能的影响。结果见表1。

表1：含异常值模型的参数以及检测效果

	系数阵	均值向量	协方差阵	UCL	报警次数
对照组	$\Phi=\begin{bmatrix}0.26 & 0.18\\ 0.24 & 0.36\end{bmatrix}$	$\mu=\begin{bmatrix}0.03\\ 0.01\end{bmatrix}$	$\Sigma=\begin{bmatrix}1.09 & 0.58\\ 0.58 & 1.08\end{bmatrix}$	2.22	4(0)
实验组	$\Phi=\begin{bmatrix}0.38 & -0.17\\ 0.53 & -0.24\end{bmatrix}$	$\mu=\begin{bmatrix}0.39\\ 0.58\end{bmatrix}$	$\Sigma=\begin{bmatrix}2.49 & 2.62\\ 2.62 & 4.38\end{bmatrix}$	3.23	5(10)

由表1可以看出，当时间序列中掺入10%的异常观测后，传统基于样本均值向量和协方差矩阵的估计结果出现明显偏移。由于这类统计量对异常值高度敏感，参数估计误差在VAR(1)模型拟合过程中被进一步放大，导致模型系数发生显著变化。受此影响，基于残差构造的监控统计量Z_t以及相应的控制上限均发生系统性偏移，从而削弱了控制图对过程异常的识别能力。

再比较污染前后的控制图参数，异常值一进来，上控制限（UCL）就明显抬高，传统VAR控制图在异常数据环境下的稳健性不足，从这里就能看出来。控制限被抬高，等于把报警门槛往上挪，控制图整体跟着变迟钝，过程真的发生偏移时往往不能及时报警，过程失控却未被识别的风险也就上去了。与此同时，异常观测还会干扰对过程结构的判断，把潜在的变化模式盖住，控制图在过程诊断和问题定位中的辅助作用也跟着打折扣。

污染率较高时，还有另一个问题：同一时刻常常有多个变量同时受到异常干扰。但传统VAR控制图采用

$$Z_t=\max_{1\le i\le p}|\ Z_{it}|$$

作为综合监控统计量，报警机制只指向最显著的那个失控变量，多变量同时异常的情况很难被完整反映出来。比如在本次模拟实验中，第35期观测点发生异常时，多个变量都受到了污染影响，可控制图只对其中一个变量给出报警，其余变量的同步失控风险被放过去了。放在实际工业决策中，这就可能造成信息利用不充分，甚至误判，如表2所示。

表2：检测结果（部分）

No.obs	X_1	X_2	$\lvert Z_{it}\rvert$	$\lvert z_{2t}\rvert$	Z_t
34	−0.2971	−0.2537	0.3280	0.3969	0.3969
35	2.6230 (−0.40025)	7.0736* (1.66285)	1.4141	4.1130*	4.1130*
36	0.6955	−0.9947	0.1463	0.7511	0.7511
37	1.4739	0.2576	0.6866	0.2022	0.6866

在多变量过程监控中，若控制图每次报警后均需对所有变量进行逐一排查，将显著增加时间与人工成本，尤其在高维情形下更为突出。因此，传统仅提供整体失控信号的控制图难以满足实际决策需求，亟需一种兼具稳健性且能够准确指示失控变量的多变量控制图方法。

二、稳健向量自回归控制图的构建

（一）分位数加权的计算

在通过考虑每个Z_{it}在其变量分布中的相对重要性（使用分位数作为权重）来更精细地整合多变量的信息。这种方法的核心思想是根据变量值的统计重要性赋予不同的权重，从而可能提高异常检测的灵敏度和特异性。

首先，对于观测序列矩阵的每个数据点，计算其标准化分数Z：

$$Z_{it}=\frac{X_{it}-\mu_i}{\sigma_i}$$

其中，X_{it}是第i个时刻第t个变量的观测值，μ_i是对应变量的均值，σ_i是对应变量的标准差。接着计算其在所有$|Z_{it}|$中的分位数：

$$q_{it}=F_{Z_i}\left(|Z_{it}|\right)$$

其中，F_{Z_i}是第i个变量的Z值的分布函数，由于在实际生产过程中异常值的在总过程中的占比不会超过50%，这里仅考虑变量的异常程度足够大的时候才考虑其对整体的影响，进一步减少噪声的干扰，使控制图对过程中真正的变化更为敏感。根据分位数q_{it}计算权重：

$$\omega_{it}=\begin{cases}0, q_{it}<0.5\\ q_{it}, q_{it}\ge 0.5\end{cases}$$

对于每个时刻的Z_{it}，把它的权重做归一化处理，让每个时刻的总权重恰好为1：

$$\omega_{it}^{'}=\frac{\omega_{it}}{\sum_{t=1}^{p}\omega_{it}}$$

有一种情况需要单独处理：如果某时刻所有ω_{it}都是0（也就是所有q_{it}都小于0.5），那就让每个$\omega_{it}^{'}$直接取$\frac{1}{p}$。

接着算每个时刻的加权Z_t值，权重就用归一化后：

$$Z_t=\sum_{t=1}^{p}\omega_{it}^{'}|Z_{it}|$$

再把优化后的Z_t值的$(1-\alpha)$分位数拿来当上控制限：

$$UCL=F_{Z_t}^{-1}(1-\alpha)$$

其中a是置信水平，$F_{Z_t}^{-1}$是Z_t分布的逆函数。

（二）稳健打点值的选择

设计多变量自相关控制图时，传统做法一般靠样本均值和协方差矩阵来估计过程的中心和散布。本文换一个思路，从仿射不变性和高崩溃点两个角度来挑稳健估计量。仿射不变性确保的是估计结果别受变量线性变换和尺度差异的影响；高崩溃点确保的是在存在一定比例异常值的情况下，估计量还能不能稳住。基于上述准则，本文选了RMCD和RMVE估计量，用于后面稳健控制图的搭建。

RMCD估计量的思路是：先挑出一个协方差行列式最小的子集，再拿这个子集去估计位置和散布。简单来说，RMCD估计量

Copyright © This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

旨在寻找一个具有最小行列式的子样本集，从而最小化协方差矩阵的估计误差，让估计量对异常值不那么敏感。

给定样本数据集$\{\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_n\}$，RMCD 的估计分三步走：

1. 选择子集：从样本数据里挑出一个大小为 h 的子集，其中 $h \le n$，选择的标准是这个子集的协方差矩阵行列式最小。

2. 估计位置：把子集的均值拿来当稳健位置估计，记为：

$$\hat{\mu}_{\text{RMCD}} = \frac{1}{h}\sum_{i=1}^{h}\mathbf{x}_i$$

3. 估计散布：把子集的协方差矩阵拿来当稳健散布估计，记为：

$$\hat{\Sigma}_{\text{RMCD}} = \frac{1}{h-1}\sum_{i=1}^{h}(\mathbf{x}_i-\hat{\mu}_{\text{RMCD}})(\mathbf{x}_i-\hat{\mu}_{\text{RMCD}})^T$$

此外，RMVE 估计量通过最小化覆盖数据的椭球体体积来进行散布的估计。该方法选择一个适当的子集，其余估计步骤和 RMCD 估计量一致，在此不再赘述。

通过计算 RMCD 和 RMVE 的稳健估计量，可得出下面是基于 RMCD 和 RMVE 对应的 Z_t^R：

$$Z_{t-RMCD}^R = \sum_{t=1}^{p}\omega_{it}^{'}Z_{it-RMCD}^R \quad (1)$$

$$Z_{t-RMVE}^R = \sum_{t=1}^{p}\omega_{it}^{'}Z_{it-RMVE}^R \quad (2)$$

以上为稳健控制图中稳健打点值部分的优化步骤。

（三）控制限的优化计算

本文通过蒙特卡洛模拟以获得不同参数下对应的控制限 UCL。从标准多变量正态分布 $MVN(0,I_p)$ 中生成了 n=2000 个大小为 m 的样本，维度为 p。由于 Z_{t-RMCD}^R 和 Z_{t-RMVE}^R 统计量具有不变性，这些限值适用于任何 μ 和 Σ 的值。本文使用崩溃值为 $\gamma=0.50$ 的加权 MCD/MVE 估计量（$\bar{x}_{RMCD}, S_{RMCD}, \bar{x}_{RMVE}$ 和 S_{RMVE}），按照公式（1）（2）计算观测样本数据集中每个观测值的 Z_{t-RMCD}^R 和 Z_{t-RMVE}^R 统计量，并记录每个大小为 m 的数据集中达到的最大值。通过计算 Z_{t-RMCD}^R 和 Z_{t-RMVE}^R 的逆函数的最大值的经验分布，确定了 $(1-\alpha)100\%$ 的分位数。本文使用了由 Torodov 编写的 R 函数“CovMcd()”在“rrcov”包中计算 RMCD/RMVE 估计量。

本文构建了在 γ=0.50 的条件下，样本量 (m=[30(1)50,55(5.100, 110(10)200])，维度 (p=(2,3…10))，以及 (a=(0.05,0.01,0.005)) 时的 Z_{t-RMCD}^R 和 Z_{t-RMVE}^R 的经验分布。

$$f_{p,\alpha,\gamma,m} = a_{1,(p,\alpha,\gamma)} + \frac{a_{2,(p,\alpha,\gamma)}}{m^{a_{3,(p,\alpha,\gamma)}}} \quad (3)$$

表3为不同参数对应下 UCL 控制限拟合的参数估计，为节约空间仅展示常用参数下的估计值。

表3：Z_{t-RMCD}^R 和 Z_{t-RMVE}^R 控制图的模型参数 $a_{1,(p,\alpha,\gamma)}$，$a_{2,(p,\alpha,\gamma)}$ 和 $a_{3,(p,\alpha,\gamma)}$ 的估计

p	a=0.05				a=0.01			
	$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$	R^2	$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$	R^2
				Z_{t-RMCD}^R				
2	2.067	15.934	1.128	0.979	2.582	12.543	1.091	0.958
4	2.007	157.032	1.580	0.987	2.413	168.087	1.566	0.978
6	1.901	571.627	1.816	0.990	2.222	656.758	1.806	0.988
8	1.781	533.681	1.700	0.988	2.022	622.341	1.683	0.987
10	1.680	459.048	1.591	0.992	1.880	616.460	1.611	0.990
				Z_{t-RMVE}^R				
2	2.179	34.903	1.290	0.992	2.710	20.057	1.146	0.984
4	2.084	226.930	1.786	0.997	2.519	530.152	2.020	0.992
6	1.971	162.440	1.697	0.997	2.312	223.008	1.759	0.995
8	1.881	162.676	1.618	0.997	2.153	173.112	1.596	0.997
10	1.795	115.033	1.441	0.999	2.028	156.129	1.478	0.999

（四）稳健向量自回归控制图的优化步骤

构建稳健多变量自相关控制图的具体步骤如下：

阶段 1

1. 确定样本量 n、变量数 p 以及置信水平 $1-\alpha$。

2. 以明确的周期间隔收集阶段 I 数据 $\{x_1,\cdots,x_n\}$。

3. 根据样本规模参数选调合适的稳健统计量（下以 RMCD 为例），使用阶段 I 数据计算 RMCD 的估计值 x_{RMCD} 和 S_{RMCD}，并选择崩溃点 $1-\gamma=0.5$。

4. 根据期望的 a 和 p 值，从表格 3 中选择最小二乘法估计参数 $a_{1,(p,\alpha,\gamma)}$，$a_{2,(p,\alpha,\gamma)}$ 和 $a_{3,(p,\alpha,\gamma)}$ 然后使用公式 (3) 计算控制限。

阶段 2

5. 根据公式（3-14，15）计算每个新观测值的 Z_t^R，并将其绘制在已在阶段 I 中（步骤 4）确定的控制限的控制图上。

6. 对控制图进行动态监测，及时查找失控点确定农长变量。如有必要，对过程进行诊断。

通过以上步骤，即可有效地实施稳健多变量自相关控制图来监控过程中的多元数据。

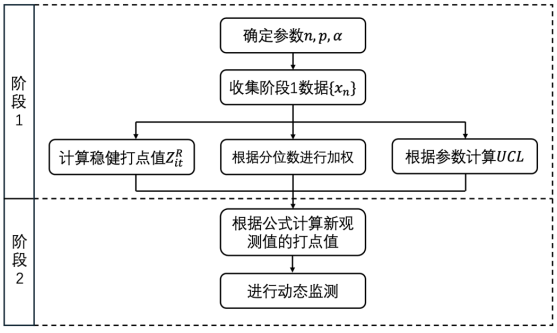


图1：稳健多变量自相关控制图建模流程图

三、模拟与实证分析

（一）模拟分析

为了说明稳健 VAR 控制图在实际生产过程中有良好的检测效力，本文将模拟一个含污染值的 VAR(1) 过程，构建含污染样本的目的是模拟现实世界数据中可能遇到的异常情况，这些异常值可能由不同来源引起，如测量误差、数据录入错误或极端事件等。通过在模拟的 VAR 过程中人为添加污染值，我们可以评估控制图

 Copyright © This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

在实际应用中面对异常值时的表现和稳健性。

模拟实验这样安排。先定义基本的 VAR(1) 模型，其中 X_t 表示在时间 t 的多变量观测值向量，Φ 为系数矩阵，ε_t 为误差项：

$$X_t = \Phi X_{t-1} + \varepsilon_t$$

这里假定 ε_t 为具有协方差矩阵 Σ 的多元正态分布，即 $\varepsilon_t \sim N(0,\Sigma)$。

污染比例设为 a，一般取 $0<\alpha<1$，对于发生异常值的时刻 t，样本值为：

$$X_t = X_t + \delta_t$$

其中，δ_t 是叠加到观测值 Y_t 上的污染向量，用来模拟异常情况下的数据点。实际操作中，δ_t 可以从均值（μ_δ）和协方差矩阵（Σ_δ）都与原分布明显不同的多元正态分布中抽取，以体现异常值的特征，即 $\delta_t \sim N(\mu_\delta,\Sigma_\delta)$。这样一来，污染值会让被选中的观测值明显偏离原有数据分布，现实中可能遇到的异常情况也就被还原出来了。

下面展示的模拟演示，均基于样本量 $n=500$，VAR(1) 模型系数 $\Sigma=\begin{bmatrix}0.5 & 0.2\\ 0.1 & 0.4\end{bmatrix}$，均值向量 $\mu=\begin{bmatrix}0\\0\end{bmatrix}$ 的条件进行，结果如图2和表3所示。

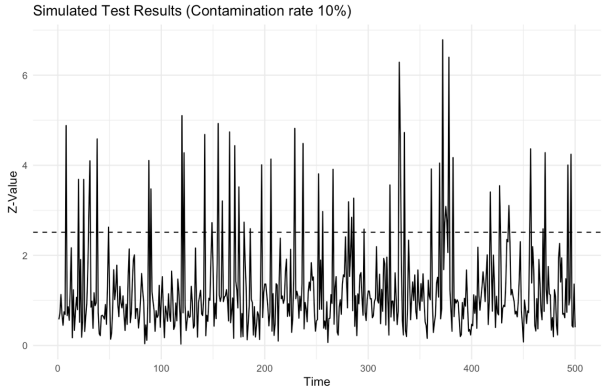


图2：在污染率为10% 条件下的模拟检测效果

表4展示了不同控制图在不同污染率下的检测效果。监控多变量时间序列数据中的异常活动时，3Sigma 控制图靠设定统计阈值来识别偏差，残差 VAR 控制图则检查 VAR 模型的残差，这两种都是生产过程中常见的控制图。下面把它们与本文提出的的稳健 VAR 控制图放在一起对比，看看在不同程度的数据污染下，谁识别异常的能力更强，进而确定复杂数据环境下更管用的监测工具。

表4：不同控制图在不同污染率下的检测效果

污染率	$\alpha=5\%$			$\alpha=15\%$			$\alpha=30\%$		
	3Sigma	残差 VAR	稳健 VAR	3Sigma	残差 VAR	稳健 VAR	3Sigma	残差 VAR	稳健 VAR
报警个数	190	9	39	178	11	72	234	2	112
准确率	67.00%	96.80%	96.80%	78.20%	87.20%	96.60%	70.00%	70.40%	90.80%
查准率	13.16%	100.00%	61.54%	40.44%	100.00%	90.28%	50.00%	100.00%	96.43%
查全率	100.00%	36.00%	96.00%	96.00%	14.67%	86.67%	78.00%	1.33%	72.00%
F1分数	0.23	0.53	0.75	0.57	0.26	0.88	0.61	0.03	0.82

把不同污染率下三种控制图的表现摆在一起看，残差 VAR 控制图对异常值的检测偏保守，查全率因此远低于稳健 VAR 控制图，到了高污染率（$\alpha=30\%$）时，直接检测失效。

传统的3Sigma 在低污染率下能把真实异常值全部揪出来，可误报也跟着上来了，查准率并不好看。污染率一往上走，它的准确率和查准率还在往下掉，这一点也就更明显。查全率是保住了，但实际用起来，频繁误报既浪费资源，也容易让人对报警麻木。

稳健 VAR 控制图的表现就不一样了，污染率怎么变它都比较稳。污染率升上去，它照样能把异常值认出来，报警不多，漏报也少，准确率和查准率因此一直在线。更值得一提的是，污染率往上走，稳健 VAR 控制图的性能并没有明显往下掉，崩溃点和稳健性都撑得住。它的 F1分数在所有污染条件下都保持在高位，说明它在查准和查全之间找到了平衡，这一点在实际应用中很要紧。

接下来，研究把目光转向维度。在中等污染率（α=15%）下逐步加维度，看看控制图面对更复杂的数据结构还稳不稳，见表5。

表5：污染率 $\alpha=15\%$ 下不同维度的检测效果

	p=2	p=3	p=4	p=5	p=6
报警个数	72	75	60	62	66
准确率	96.60%	99.20%	96.60%	97.00%	97.00%
查准率	90.28%	97.33%	98.33%	98.39%	95.45%
查全率	86.67%	97.33%	78.66%	81.33%	84.00%
F1分数	0.88	0.97	0.87	0.89	0.89

表5显示，在中等污染率（$\alpha=15\%$）下，稳健 VAR 控制图的性能随维度增加基本能稳住。报警个数在不同维度 (p=2至 p=6) 之间有些起伏，但整体还在相对稳定的范围内，没有出现维度越高报警越多或越少的趋势。控制图在标识潜在异常时的敏感度，并没有因为维度增加而被明显拉低。

更值得一提的是，各维度下准确率都超过了95%，这个数字说明控制图能把正常点和异常点分得比较清楚，即便数据里掺了污染，也还能保持可靠。查准率和查全率同样稳当，没有随着维度增加而明显下滑，说明控制图在不牺牲太多敏感性的前提下，仍能维持较高精确度。所有维度中 F1分数都在0.87以上，这些结果说明控制图在各个维度上都能保持住自己的检测性能。

为了进一步看稳健 VAR 控制图在不同维度下的检测情况，下面用 F1分数来综合刻画它的检测效力，对不同维度做多次模拟实验，并画出小提琴图，见图3：

从图3可以看出，维度增加后，稳健 VAR 控制图的 F1分数中位数在大多数维度上都维持在较高水平，分布也大多比较集中，说明高维情况下该控制图的表现相对稳定。不过，$p=5,6$ 时出现了少量的离群点，F1分数的整体分布范围也更宽。这一点值得留意：维度再往上走，模型复杂度跟着涨，稳健估计量本身也会开始吃力，性能波动因此被放大。

Copyright © This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

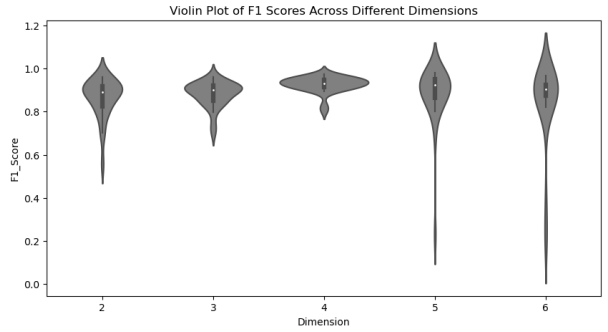


图3：不同维度下 F1 值的小提琴图

综上，不同维度和污染率下的实验结果表明，稳健 VAR 控制图面对复杂数据时稳得住。即便污染率高达30%，准确率仍保持在90% 以上，报警个数也只是适度增加，说明它对异常的识别既敏感又准确。F1 分数在不同维度上的稳定性说明，该控制图能在查准率和查全率之间取得平衡，检测性能也因此较为可靠。

（二）实证分析

模拟完成，接下来把方法放到真实数据测试一下。

本研究使用了广州市2022年的每日气象数据，重点分析空气质量相关的六个污染物指标，包括 PM2.5、PM10、SO2、NO2、O3（8小时均值）和 CO。数据集共365条记录，每条代表一天的污染物浓度和相应 AQI 级别。AQI 级别是评估空气质量的标准，也是衡量空气污染程度和对公众健康影响的关键指标。在该数据集中，AQI 级别大于等于3意味着空气质量较差，需要做异常值分析，找出可能由环境或人为因素导致的空气质量异常。

应用稳健 VAR 控制图方法后，我们在广州市2022年的气象数据中识别出若干异常值。这些异常值对应的日子主要集中在特定时间内，可能与季节变化、人为活动增加或特殊环境事件有关，见图4。

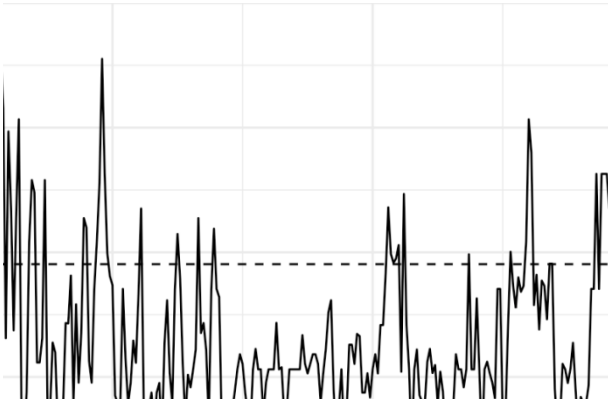


图4：广州市2022年空气污染物稳健控制图

为验证异常值检测的准确性，我们把检测到的异常值与 AQI 级别做了对比。当检测到的异常值对应的 AQI 级别大于等于3（轻度污染）时，就认为这次异常值检测是准确的。

图5是本实证结果的前100期可视化的对比图，其中浅灰色色块代表正常值，未检测出异常情况，黑色为发生异常的变量，颜色的深浅代表着影响程度的强弱。不难发现，在处理多变量问题时，由于数据的复杂度增加，传统的 VAR 控制图已经不能准确的识别失控过程，使得检测结果可信度降低，反观稳健 VAR 控制图，在处理高维数据时依然能表现稳定的检测效力；更值得注意的是，稳健 VAR 控制图还提供了不同变量在异常时刻的影响程度可视化，根据各时刻加权 Z_t 的系数可以绘制以上影响程度色块图，这里不但指出了异常时刻中影响程度最大的变量，还一并将发生变异的其余变量通过不同深浅的颜色标注出来，为决策者提供了准确的异常变量信息，以便决策者后续的判断和调整。

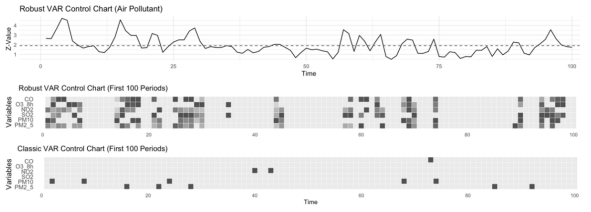


图5：传统 VAR 控制图和稳健 VAR 控制图检测结果（部分）的可视化

根据广州市2022年的大气数据，基于本文提出的稳健 VAR 控制图检测出异常值检测的准确率为83.01%。这表明，稳健 VAR 控制图方法在广州市2022年的气象数据中表现出了较高的准确性，能够有效地识别空气质量的异常变化。

四、结束语

在本研究中，我们对面临数据污染时多变量时间序列监控的挑战进行了系统的探索，详细评估了稳健 VAR 控制图的性能。模拟实验展示了在不同污染率下，稳健 VAR 控制图相较于传统 3Sigma 和残差 VAR 控制图，无论在低还是高污染水平，均能有效地识别异常值。

特别是在中等至高污染率下，稳健 VAR 控制图展现了其优越的稳定性和较高的崩溃点，整体准确率高达90% 以上。其 F1 分数普遍高于0.8，彰显了即使在数据复杂性升高的情形下，该方法依然能够提供精准的警报，最小化误判，并且在高维数据上表现出稳定的检测效力。并且可以为决策者提供发生异常时刻各影响变量的程度可视化，能够帮助决策者找到并分析异常值的相关影响变量，提升监测精度的同时提高了检测效率避免了不必要的浪费。

综上，通过在多种污染场景下的强表现，本研究提出的稳健 VAR 控制图证明了自己在多变量时间序列数据监控领域的实用性和可靠性。其稳健性和高准确率使之成为适用于当今数据密集型产业的理想工具，尤其是在存在数据质量问题的实际应用中。未来的研究将建立在这一基础之上，进一步扩展控制图在多元数据环境中的应用，以不断增强其在过程监控中的效率和效果。

参考文献

[1]Centofanti, Fabio, Antonio Lepore, Alessandra Menafoglio, Biagio Palumbo, and Simone Vantini. Functional Regression Control Chart[J]. Technometrics,2021,63(3):281 - 94.

 Copyright © This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

[2]Elisa C,E. R L. Robust multivariate control chart based on shrinkage for individual observations[J].Journal of Quality Technology,2022,54(4):415-440.

[3]Bayati M ,Nasrollahzadeh S ,Moghadam B M . Robust Monitoring of Multivariate Location Using a New Nonparametric Depth - Based Control Chart[J].Journal of Applied Mathematics,2026,2026(1):1-13.

[4]Pulido D J A ,Franco C E A ,G ó mez T G V . Nonparametric multivariate control chart with guaranteed in-control performance and cautious learning[J].Computers &Industrial Engineering,2026,218:112124.

[5] 王志坚 , 郑健松 . 稳健 Shewhart 控制图的构建及应用 [J]. 统计与决策 ,2018,34 (24): 39-42.

[6]AliH T ,SedeeqS B ,SalehM D , et al. Robust multivariate quality control charts for enhanced variability monitoring[J].Quality and Reliability Engineering Internation al,2023,40(3):1369-1381.

[7]S V ,Sebastian N ,TA S . Robust Bivariate Control Chart for Process Monitoring with Rational Subgroups[J].Quality and Reliability Engineering Internation al,2025,41(7):3232-3243.

[8] 薛丽 , 李聪凯 , 贾元忠 , 等 .VSI 多变量自相关过程 VAR 控制图经济设计 [J]. 数理统计与管理 ,2023,42(04):626-636.

[9]Capezza C, Centofanti F, Lepore A, etal. Functional regression control chart for monitoring ship CO2_2 emissions[J].Quality and Reliability Engineering Internation al,2022,38(3):1519-1537.

[10]Will M ,Mark R . SampleQC: robust multivariate, multi-cell type, multi-sample quality control for single-cell data.[J].Genome biology,2023,24(1):1-22.

[11] 雷彬文 , 龚立雄 . 多变量控制图及过程能力指数在生产中的应用 [J]. 湖北工业大学学报 ,2023,38 (01):5-10.

[12]Kalgonda, A. A., and S. R. Kulkarni. Multivariate Quality Control Chart for Autocorrelated Processes[J].Journal of Applied Statistics, 2004,31 (3):317 - 327.

[13]Hawkins, Douglas M. Multivariate Quality Control Based on Regression-Adiusted Variables[J].Technometrics,1991,33(1):61 - 75.

Copyright © This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License