Statistics and Application 统计学与应用, 2026, 15(1), 1-7
Published Online January 2026 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2026.151001

Hans汉斯

基于分类算法的变量选择控制图

徐亚萍，訾雪旻*

天津职业技术师范大学理学院，天津

收稿日期：2025年12月1日；录用日期：2025年12月22日；发布日期：2026年1月4日

摘　要

高维稀疏过程监控的理论方法需通过真实场景验证才能落地，针对生物信息学、工业生产中数据分布非理想、变量相关性复杂、噪声干扰显著的实际监控痛点，本文基于提出的L0-L2组合正则化变量选择理论，通过L0-L2组合正则化不仅能够选择变量并收缩系数，还能高效处理相关特征，精准识别异常变量，同时利用逻辑回归模型感知特定方向的偏移，并采用极大值函数将二者动态融合，形成一个具有方向自适应的监控统计量。它是一种新的变量选择控制图(LQSVS)，结合了分类算法来解决高维、稀疏分类问题，主要聚焦理论创新与模拟验证。现开展真实数据应用验证与优化研究，以UCI大肠杆菌蛋白质数据集为研究对象，首先针对真实数据特性完成预处理，并采用Bootstrap重抽样技术优化控制限计算；其次通过控制变量实验确定最优参数；最终在平均运行长度基准$ARL_0 = 200$下，验证该方法对失控(OC)数据的平均检测延迟ARL_1低至1.68，结果显著优于传统控制图。实验结果表明，所提方法可有效解决真实高维数据中“稀疏偏移检测灵敏度低、参数适配难”的问题，为蛋白质定位监控、工业多变量过程诊断等场景提供了实用工具。

关键词

分类算法，统计过程控制，变量选择，大肠杆菌蛋白质监控，高维稀疏过程

Variable Selection Control Chart Based on Classification Algorithm

Yaping Xu, Xuemin Zi*

School of Science, Tianjin University of Technology and Education, Tianjin

Received: December 1, 2025; accepted: December 22, 2025; published: January 4, 2026

Abstract

The theoretical methods for high-dimensional sparse process monitoring can only be put into practical application after validation in real scenarios. Aiming to address practical monitoring pain points,

*通讯作者。

文章引用: 徐亚萍, 訾雪旻. 基于分类算法的变量选择控制图[J]. 统计学与应用, 2026, 15(1): 1-7.
DOI: 10.12677/sa.2026.151001

such as non-ideal data distribution, complex variable correlation, and significant noise interference, in bioinformatics and industrial production, this paper is based on the proposed L0-L2 combined regularization variable selection theory. The L0-L2 combined regularization can not only select variables and shrink coefficients, but also efficiently handle correlated features and accurately identify abnormal variables. Meanwhile, the logistic regression model is used to sense shifts in specific directions, and the maximum function is adopted to dynamically integrate the two, forming a direction-adaptive monitoring statistic. It is a new variable selection control chart (LQSVS), which combines classification algorithms to solve high-dimensional and sparse classification problems, focusing mainly on theoretical innovation and simulation verification. Now, research on real-data application validation and optimization is carried out, taking the UCI *E. coli* protein dataset as the research object. Firstly, preprocessing is completed according to the characteristics of real data, and the Bootstrap resampling technique is used to optimize the calculation of control limits. Secondly, the optimal parameters are determined through controlled variable experiments. Finally, under the benchmark of in-control average run length (ARL_0) = 200, it is verified that the average run length for out-of-control (OC) data (ARL_1) of this method is as low as 1.68, which is significantly better than that of traditional control charts. The experimental results show that the proposed method can effectively solve the problems of "low sensitivity to sparse shift detection and difficult parameter adaptation" in real high-dimensional data, and provide a practical tool for scenarios such as protein localization monitoring and industrial multivariate process diagnosis.

Keywords

Classification Algorithm, Statistical Process Control (SPC), Variable Selection, *E. coli* Protein Monitoring, High-Dimensional Sparse Process

1. 引言

高维稀疏过程监控是现代工业生产、生物信息学等领域的核心需求——例如化工多变量生产过程中仅 2~3 个关键变量偏移就可能引发质量事故，大肠杆菌蛋白质定位研究中需从高维数据中快速识别异常定位信号。传统 Hotelling T^2 控制图对所有变量全局监控，在真实高维数据中因“维度诅咒”导致稀疏偏移检测灵敏度极低[1]；LASSO-VS 控制图虽通过变量选择优化，但仅依赖在控(IC)数据构建模型，无法利用历史失控(OC)数据中的故障模式信息，在真实场景中易受变量相关性、随机噪声干扰；概率分类(PoC)图虽融入 OC 数据，却对预设偏移方向过度依赖，难以适配真实数据中偏移方向不确定的复杂情况。当前面临理论与实际脱节的问题。

为解决高维稀疏监控的理论创新问题，基于 L0-L2 组合正则化[2]的变量选择控制图(LQSVS)，通过数值模拟验证了核心优势。L0 范数实现精准稀疏变量筛选，避免传统 LASSO 在高相关数据中随机选择单一变量的缺陷[3] [4]；L2 范数收缩非零系数，降低噪声对模型的干扰；变量选择统计量和分类边界距离统计量双统计量融合机制，兼顾特定方向与非特定方向的偏移检测。但该研究仅基于理想高相关模拟数据，比如预设正态分布、固定相关性系数、低随机噪声。未涉及真实数据可能存在部分变量线性相关、真实数据分布常偏离理想正态、理论参数在真实场景中的最优取值等问题，需针对性预处理以匹配理论假设。

本文主要以 UCI 大肠杆菌蛋白质数据集为研究对象，该数据集具有高维属性、强相关性、分布复杂性。原始含 7 个变量，经预处理后保留 5 个关键变量，符合高维监控场景，且部分样本偏离正态分布。下面主要针对真实数据特性优化 LQSVS 控制图的技术细节，使高维稀疏监控理论落实到实际应用。

2. 基于分类算法的 L0-L2 变量选择控制图

基于逻辑回归模型的分类边界[5]距离统计量定义为：

$$D_c\left(x_t\right)=\mathrm{e}^{d_c\left(x_t\right)} \tag{1}$$

其中 $d_c\left(x_t\right)=\beta_0^*+\sum_{i=1}^{p}\beta_i^* x_i$ 测量观测值与边界之间的距离，$\beta_i\left(i=0,1,2,\cdots,p\right)$ 是逻辑模型的回归系数[6]。

变量选择通过以下正则化目标函数实现：

$$\text{minimize}\ \lambda\sum_{j=1}^{p}I\left(\beta_j\neq 0\right)+\gamma\sum_{j=1}^{p}\left|\beta_j\right|^2 \tag{2}$$

其中，β_j 是模型系数，$I(\cdot)$ 是示性函数，当 $\beta_j\neq 0$ 时取值为 1，否则为 0。λ 和 γ 是正则化参数，控制着 L0 和 L2 正则化的强度。L0-L2 组合正则化方法是一种先进的变量选择[7]技术，旨在解决高维数据中的稀疏性问题。通过优化上述目标函数，可以实现在保持模型稀疏性的同时，对非零系数进行适当的收缩。

在监控问题中，转化为带 L0-L2 惩罚的马氏距离最小化问题，将含惩罚项的目标函数定义为：

$$S\left(x_t\right)^{(1)}=\min_{\mu\in\Omega_1}\left(x_t-\mu\right)^T\Sigma^{-1}\left(x_t-\mu\right)+\lambda R+\gamma\sum_{j=1}^{p}\left|\mu_j\right|^2 \tag{3}$$

其中 $R=I\left(\left|\mu_j\right|\neq 0\right)$ 表示筛选出的异常变量数量，λ 和 γ 分别为 L0、L2 正则化参数；通过乔列斯基分解 $\left(\Sigma^{-1}=Y^TY\right)$ 将式(3)转化为最小二乘形式

$$S\left(x_t\right)^{(2)}=\min_{\mu\in\Omega_1}\left(z_t-Y\mu\right)^T\left(z_t-Y\mu\right)+\lambda R+\gamma\sum_{j=1}^{p}\left|\mu_j\right|^2 \tag{4}$$

此时 $z_t=Yx_t$ 是预测变量，Y 为标准化观测矩阵，该转化使马氏距离最小化问题更易应用 L0-L2 正则化。当新观测值到来时，通过求解式(4)可得到稀疏均值估计 μ^*，进而构建变量选择统计量：

$$d_{vs}\left(x_t\right)=x_t^T\Sigma^{-1}\mu^*\left(x_t\middle|R\right) \tag{5}$$

该统计量用样本协方差矩阵 S 替代总体协方差矩阵 Σ，兼具稀疏性与距离度量特性。自然地，结合公式(1)和(5)这两种统计量，最终监控统计量通过极大值函数融合两类统计量：

$$\Lambda_{svs}\left(x_t\right)=\max\left(d_{vs}\left(x_t\right),k\cdot D_c\left(x_t\right)\right) \tag{6}$$

其中 k 为尺度因子，用于统一量纲。该统计量具备方向自适应性，可同时应对特定方向与非特定方向的稀疏偏移。

本文采用的 L0-L2 组合正则化变量选择方法及其在监控统计量中的融合机制(已投稿)，上文仅列出关键公式与统计量定义，详细理论推导与模型构建过程可参考前述研究。

多变量统计过程控制(MSPC)旨在监测高维数据的均值向量 μ 是否发生偏移[8]，即检测均值向量的变化。在新型控制图应用前，需突破传统分布假设的局限确定控制限。通过自适应学习机制，构建基于 Bootstrap 重抽样控制限体系。在本研究中，变量选择算法估计的 μ^* 是变化的，监控统计量 $\Lambda_{svs}\left(\mathrm{x_t}\right)$ 的分布很难确定，通过 Bootstrap 重抽样方法渐近地获得分布。

3. 实际应用

3.1. 数据集描述

为了在实际例子中证明提出方法的有效性，本节使用了来自加州大学欧文分校(UCI)机器学习库

(http://www.example.com)维护的真实的大肠杆菌蛋白质数据集[9]，该数据集最初是为了预测这些蛋白质的定位位点而创建的，可用于测试分类算法。数据集中有 336 个样本。对于每个样本，七个属性变量可视为观测值，而蛋白质定位位点可视为预测变量。根据不同的定位位点，数据集可分为八组 cp、im，imS，imL，imU、om，omL，pp (143、77、2、2、35、20、5 和 52)。具体信息可在引用[9]论文中找到，它描述了一种分类模型，它可以被看作是一个概率模拟决策树或贝叶斯网络的限制形式。在这种情况下，这些点渐近满足正态性假设，cp 可被视为控制内(IC)观测值共计 143 个。部分 im，imS，imL，imU 点被视为预定义失控(OC)观测值共计 116 个，而其他部分点 om，omL，pp 被视为未定义 OC 观测值共计 77 个。原始蛋白质数据含 7 个变量，其中两个变量线性相关，故通过方差膨胀因子(VIF)筛选，剔除这两个变量，最终保留 5 个独立关键变量，确保理论框架与真实数据结构匹配。

为了评估控制图的性能，通过对真实数据集展开研究。数据观测值渐近满足多元正态分布，即 $x_t \sim N(\mu_0, \Sigma)$，其中 $\Sigma = (\sigma_{ij})$ 是一个常数矩阵。考虑变量间相关性，数据模拟中是设定 ρ 值，即 $\sigma_{ij} = \rho^{|i-j|}$，现研究真实数据，先对真实数据做预处理，求其均值和方差，去除线性相关数据(两列)，研究 336 行 5 维的数据，对过程所处控制状态时求其均值向量 μ_0；处于失控状态时，考虑两种情况：预定义失控(OC)数据和未定义的 OC 数据。在实验中，首先在数据集中有 336 个样本，用训练数据集中的 243 个观测值，包括 143 个正常控制(IC)数据和 100 个失控(OC)数据(其中 50 个预定义，50 个未定义)，一起构建逻辑回归模型。此外，这 143 个正常控制的 IC 观测值也用来通过变量选择算法 L0-L2 估计 IC 数据的中心。再用 143 个 IC 数据点，采用 Bootstrap 重抽样方法，从 IC 数据中有放回地抽取 40 个观测值作为一批次，计算每个观测值的监控统计量，重复 1000 批次获得 40000 个统计量构建经验分布，得到 IC 统计量分布，再用 99.5%分位数法得到控制限和可控条件(ARL_0)下的实际平均运行长度(ARL)。最后基于剩下的 93 个 OC 数据进行检验，计算失控条件下的 ARL_1。希望 ARL_1 足够小，以便能快速检测到过程的变化，也就是说 ARL_1 越小，控制图越有效。

3.2. 基于分类算法的 L0-L2 变量选择模型中进行参数估计

仅针对大肠杆菌蛋白质真实数据的特性，对关键参数进行适配调整。在实施过程监控前，需确定两个关键参数：R 和 K。R 表示用于监控的变量筛选数量，可通过先验知识或变量选择标准(如 AIC、BIC 和交叉验证)确定，实践中因变量同时变化较少而易于确定。K 为调整参数，用于平衡从过程中提取的 IC 和 OC 信息，其取值范围在(0, 1)内，通常选 0.5 以平等地利用两者信息，但也会根据具体情况调整。这两个参数对 LQSVS 图的性能至关重要，R 和 K 的合理选择能提升控制图的稳健性。

在本文中 R 是被 L0Learn 中正则化参数 λ 和正则化参数 γ 同时控制的，其中 λ 控制 L0 正则化(变量选择的稀疏性)，值越大，选择的变量越少。γ 控制 L2 正则化(系数的收缩强度)，用于平衡模型，防止过拟合。通过设计实验测试了和模拟实验已验证 $\gamma \in \{10, 0.0316, 0.0001\}$对 ARL_1影响极小，故选择固定 γ = 0.0001，在抑制过拟合的同时保留对异常信号的敏感性，选择不同的 λ 去测试(即对应不同的变量筛选数量 R)。针对过程所处控制状态与失控状态，分别求出不同模式下观测值的分布。在应用所提出的控制图之前，确定参数 k 和 R，根据实际情况选择 $k = 1$ 并设定 $R = 1, 2, 3$ 用于比较，不过因 lambda 和 gamma 的同时控制导致部分维度的 R 有缺失，缺失源于 L0Learn 的 λ 调参限制，即当 λ 过大时，某些维度无变量被选中(R = 0)导致该组合无效。

3.3. 实验结果

为评估所提出控制图的性能，本文采用了平均运行长度(ARL)指标。所有控制图的 ARL_1 计算均基于此场景下期望 ARL_0 = 200 的设定，较小的失控平均运行长度(ARL_1)表示控制图性能更优，见表 1 所示，

数据模拟实验结果见表 2，每组 ARL_1 的模拟均重复至少 1000 次。

Table 1. Comparison of the actual ARL_1 values with different numbers of selected variables when $ARL_0 = 200$. The minimum ARL_1 values obtained from experiments
表 1. 当 $ARL_0 = 200$ 时，真实数据 ARL_1 与不同变量筛选数量之间的比较。通过实验获得的最小 ARL_1 值

	SVS chart (L0-L2)		
	R = 1	R = 2	R = 3
ARL_1	1.680	1.760	2.090

Table 2. Comparison of ARL_1 with different shift magnitudes δ in data simulation experiments when p = 5 and $ARL_0 = 200$
表 2. 当 p = 5 而 $ARL_0 = 200$ 时，数据模拟实验中 ARL_1 与不同幅度偏移 δ 的比较

	SVS chart (L0-L2)		
	R = 1	R = 2	R = 3
$\delta = 1$	9.9826	9.7878	10.2398
$\delta = 2$	1.4774	1.5244	1.5024
$\delta = 3$	0.6940	0.6974	0.6836
$\delta = 4$	0.1796	0.3272	0.9356
$\delta = 5$	0.2934	0.7296	0

3.4. 实际案例的实验分析

从表 1 的实验结果以及与 Zhang 等(2023)提出基于分类算法的敏感变量选择(SVS)控制图文章中的基线图表进行比较来评估其性能[1]，可以得出本文提出的控制图在不同变量筛选数量下和大量实验中取得 ARL_1 值。尤其是在 R = 1 时，如表 1 其对所有 OC 数据的平均检测延迟(ARL_1)仅为 1.68，这意味着控制图平均在不到 2 个样本内就能探测到过程失控，快速发现异常定位，减少实验成本。不同 R 时，真实数据的 ARL_1 略高于模拟数据(p = 5 时幅度偏移 $\delta \geq 2$ 时的 ARL_1)，原因是真实数据存在随机噪声与分布偏差，验证了理论方法的抗干扰能力。

变量选择稀疏度 R 的影响，实验结果表明，当 R = 1 时，本文提出的控制图的检测性能最优，随着 R 增大至 2 和 3，ARL_1 值逐渐升高。这一现象的原因在于真实数据中的异常很可能是由少数关键变量(甚至单个变量)的偏移引起的。见图 1，R = 1 的设置迫使模型聚焦于最显著的一个异常变量，从而获得了最尖锐的检测信号。当 R 增大时，更多变量被纳入模型，虽然增强了稳健性，但也可能引入不相关的噪声变量，稀释了对核心异常变量的监控力度，导致检测灵敏度略有下降。为验证结果的可信度，对不同方法运行 1000 次模拟得到的 ARL_1 序列进行了双样本 t 检验。检验结果表明，R = 1 的 ARL_1 均值具有较强的显著性，即具有统计学上的显著差异。也表明在生物高维数据监控中，聚焦单一关键变量的稀疏选择策略更有效(如蛋白质定位的核心影响因子仅 1~2 个)，为实际应用提供了参数选择依据。同时相较于传统方法，可大幅缩短异常响应时间，减少无效实验成本。

综上所述，实验分析充分证明将变量选择与分类算法相结合的自适应机制的有效性，本文提出的控制图通过创新的自适应双统计量融合机制，克服了之前方法的单一局限性，在真实高维过程监控任务中提供了一种更快速、更灵敏的解决方案。

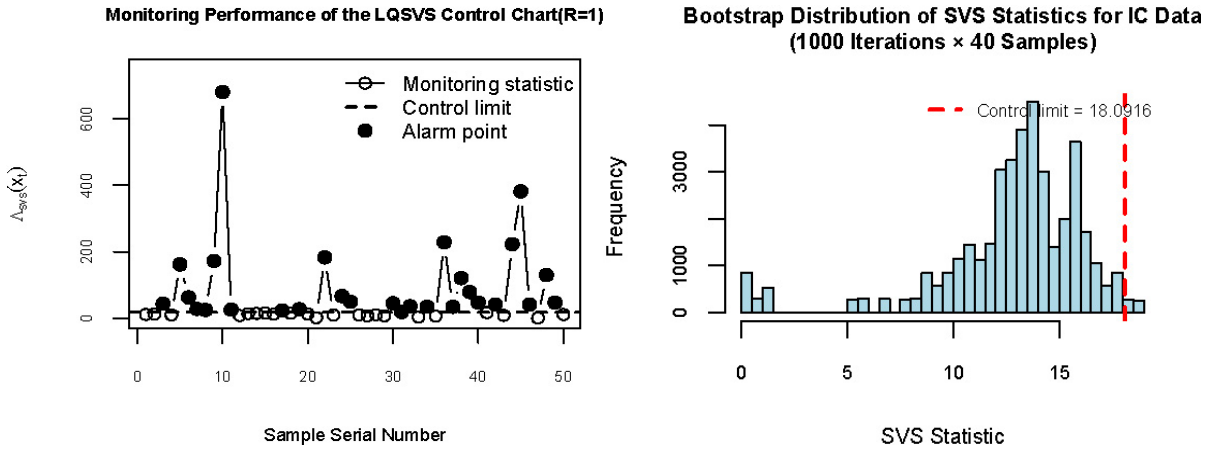

Figure 1. Monitoring performance of the LQSVS control chart and Bootstrap distribution of the SVS statistic for its in-control data when sparsity R = 1
图 1. 当稀疏度 R = 1 时 LQSVS 控制图监控和其 IC 数据 SVS 统计量 Bootstrap 分布

4. 结论

本文基于新提出的一种新颖的分类算法和变量选择相融合的多变量统计过程控制方法，通过大肠杆菌蛋白质真实数据的实验和数据模拟实验，实现了 1.68 的低 ARL_1，显著优于传统控制图，该结果验证了该方法在高维强相关性的实际场景中的实用性。结果表明，SVS (L0-L2)控制图具有更高的灵敏度，解决了真实数据中分布偏离、噪声干扰、变量冗余带来的监控难题。当前该算法在数据呈正态分布的情况下表现出稳定的性能，但是此控制图的分类模块依赖于逻辑回归算法，它对非正态数据的适配性可进一步提升，而且该方法在超大规模真实数据(如 $p > 100$)中的计算效率仍需优化。未来工作计划拓展至超大规模工业真实数据，优化算法计算效率；结合支持向量机 SVM 分类模块[10]，提升非正态、大样本真实数据的检测性能；开发可视化监控界面，降低实际工程应用的门槛。

参考文献

[1] Zhang, S., Xue, L., He, Z., Liu, Y. and Xin, Z. (2023) A Sensitized Variable Selection Control Chart Based on a Classification Algorithm for Monitoring High-Dimensional Processes. *Quality and Reliability Engineering International*, **39**, 2837-2850. https://doi.org/10.1002/qre.3393

[2] Dedieu, A., Hazimeh, H. and Mazumder, R. (2021) Learning Sparse Classifiers: Continuous and Mixed Integer Optimization Perspectives. *Journal of Machine Learning Research*, **22**, 1-47.

[3] Zou, C., Jiang, W. and Tsung, F. (2011) A Lasso-Based Diagnostic Framework for Multivariate Statistical Process Control. *Technometrics*, **53**, 297-309. https://doi.org/10.1198/tech.2011.10034

[4] Zou, C. and Qiu, P. (2009) Multivariate Statistical Process Control Using Lasso. *Journal of the American Statistical Association*, **104**, 1586-1596. https://doi.org/10.1198/jasa.2009.tm08128

[5] Zhang, C., Tsung, F. and Zou, C. (2015) A General Framework for Monitoring Complex Processes with Both In-Control and Out-of-Control Information. *Computers & Industrial Engineering*, **85**, 157-168. https://doi.org/10.1016/j.cie.2015.03.007

[6] Huang, D.X. and Lu, C.T. (2023) Several Variable Selection Methods Based on Logistic Regression Model. *Popular Standardization*, **8**, 139-141.

[7] Subbiah, S.S. and Chinnappan, J. (2021) Opportunities and Challenges of Feature Selection Methods for High Dimensional Data: A Review. *Ingénierie des systèmes d information*, **26**, 67-77. https://doi.org/10.18280/isi.260107

[8] Wang, K. and Song, Z. (2024) High-Dimensional Categorical Process Monitoring: A Data Mining Approach. *IISE Transactions*, **57**, 1088-1104. https://doi.org/10.1080/24725854.2024.2399653

[9] Horton, P. and Nakai, K. (1996) A Probabilistic Classification System for Predicting the Cellular Localization Sites of

Proteins. *International Conference on Intelligent Systems for Molecular Biology*, **4**, 109-115.

[10] Landeros, A. and Lange, K. (2022) Algorithms for Sparse Support Vector Machines. *Journal of Computational and Graphical Statistics*, **32**, 1097-1108. https://doi.org/10.1080/10618600.2022.2146697