Pure Mathematics 理论数学, 2026, 16(5), 215-226
Published Online May 2026 in Hans. https://www.hanspub.org/journal/pm
https://doi.org/10.12677/pm.2026.165145

Hans汉斯

均值模型中基于稳健变量选择的多变点估计

葛明霞，董翠玲*

新疆师范大学数学科学学院，新疆 乌鲁木齐

收稿日期：2026年4月14日；录用日期：2026年5月15日；发布日期：2026年5月28日

摘　要

基于变量选择的多变点估计是目前数据研究的热点问题，而数据中异常值的存在往往会严重干扰变点估计的精度与稳定性，基于稳健变量选择的多变点估计成为该领域亟待解决的问题。文章提出一种带有指数平方损失函数(Exponential Squared Loss, ESL)的两阶段多变点估计方法，结合调节参数对惩罚力度的灵活控制，实现了对数据中异常值的有效抵御，具备良好的稳健性。数值模拟结果表明，在含有1%、5%、10%异常值的场景下，基于指数平方损失的两阶段多变点估计方法均优于累积和方法和基于平方损失函数的两阶段多变点估计方法。应用R语言自带的测井数据进行实证分析，验证了该方法的有效性。该方法为含有异常值数据的稳健变点估计提供了有效的新思路，可为相关实际应用提供重要的方法支撑与参考。

关键词

变量选择，指数平方损失函数，均值变点模型，异常值，稳健估计

Multiple Change-Point Estimation in Mean Models Based on Robust Variable Selection

Mingxia Ge, Cuiling Dong*

School Mathematical Sciences, Xinjiang Normal University, Urumqi Xinjiang

Received: April 14, 2026; accepted: May 15, 2026; published: May 28, 2026

Abstract

Multiple change-point estimation based on variable selection is a hot topic in current data research. However, the presence of outliers in data often severely interferes with the accuracy and stability of change-point estimation. Therefore, multiple change-point estimation based on robust variable selection has become an urgent problem to be solved in this field. This paper proposes a two-stage multiple change-point estimation method with the Exponential Squared Loss (ESL) function. By

文章引用: 葛明霞, 董翠玲. 均值模型中基于稳健变量选择的多变点估计[J]. 理论数学, 2026, 16(5): 215-226.
DOI: 10.12677/pm.2026.165145

flexibly controlling the penalty intensity through tuning parameters, this method effectively resists outliers in the data and exhibits excellent robustness. Numerical simulation results show that the proposed two-stage method based on exponential squared loss outperforms both the cumulative sum (CUSUM) method and the two-stage multiple change-point estimation method based on the squared loss function in scenarios with 1%, 5%, and 10% outliers. Empirical analysis is conducted using the well-log data built into R, which verifies the effectiveness of the method. This method provides a new effective idea for robust change-point estimation of data containing outliers, and can offer important methodological support and reference for relevant practical applications.

Keywords

Variable Selection, Exponential Squared Loss, Mean Change-Point Model, Outliers, Robust Estimation

1. 引言

随着现代测量技术与数据采集手段的飞速发展，均值模型作为刻画数据中心趋势演变的基础统计模型已广泛应用于地震学[1]、金融工程[2] [3]、遗传学[4]、环境监测[5] [6]等多个领域。在这些领域中，数据的均值往往并非始终保持平稳，而是会在某些未知时刻发生结构性突变(即“变点”)，例如金融市场政策调整引发的收益率均值突变、地震活动中能量释放强度的均值改变等。准确估计均值模型中的多变点，不仅是揭示数据内在演化规律的关键，更是后续统计推断、预测决策的重要前提，具有重要的研究价值与实践意义。

1954 年 Page 提出累积和方法[7] (cumulative sun, CUSUM)，将累积和思想用于过程均值的变点估计，成为工业质量控制中估计均值变点的经典方法，也是变点研究领域的标志性起点。1971 年，Hinkley 提出基于似然比检验[8] (Likelihood Ratio Test, LRT)的变点估计方法，通过比较不同假设下的似然值确定变点位置，为参数变点估计提供了理论基础。后续学者提出了非参数方法及序贯检测方法等，可参见 Csörgö 和 Horváth (1997) [9]、Wu (2005) [10]等及其参考文献。

由于变点的稀疏性，Brodsky 和 Darkhovsky (1993) [11]开创性地采用变量选择技术来估计变点。Harchaoui 和 Lévy-Ledu (2010) [12]利用全变差惩罚函数(Total Variation Penalty)研究了干扰项为白噪声的常数信号中的多变点估计问题。2016 年 Jin 等[13]提出了一种快速且精准的基于变量选择的两阶段多变点估计方法，应用平方损失函数搭配自适应最小绝对收缩与选择算子(Adaptive Least Absolute Shrinkage and Selection Operator, Adaptive Lasso)、带平滑削边绝对偏离法(Smoothly Clipped Absolute Deviation, SCAD)和极小极大凹惩罚(Minimax Concave Penalty, MCP)三种惩罚函数，对线性回归模型进行两阶段多变点估计。该方法为多变点估计提供了极具创新性的思路，其核心逻辑在于通过构造特殊设计矩阵将“多变点的估计问题”转化为“变量选择问题”。

随着信息技术的迅猛发展和大数据时代的到来，由于测量误差、仪器故障、极端事件等因素，收集到的数据中往往包含 1%~10%的异常值[14]，在网络数据、环境监测数据等复杂场景中这一比例可能更高。因此，为了抵御异常值的影响，大量学者开始寻求稳健估计方法。1981 年 Huber 关于线性回归模型提出了 M 估计[15]，为后续各类稳健参数估计及稳健变量选择方法奠定了基础。后续学者进一步拓展了

稳健估计理论，提出了 MM 估计[16]、τ 估计[17]、R 估计[18]、LMS 估计[19]、LTS 估计[20]、S 估计[21]以及 WLS 估计[22]等一系列稳健估计方法，极大丰富了稳健统计的研究体系。

受到文献[13]的启发，文章将指数平方损失函数引入均值变点模型的两阶段多变点估计方法中，提出一种“基于指数平方损失的两阶段多变点估计”方法，命名为“Two-Stage Multiple Change Point Detection-Exponential Squared Loss”(TSMCD-ESL)。第一阶段首先对数据进行分割，引入一个特殊的设计矩阵，将均值模型的变点估计问题转化为高维线性回归模型的变量选择问题，其次构造带有指数平方损失函数的目标函数对高维线性回归模型进行变量选择，快速锁定潜在变点区域；第二阶段通过拟似然比检验确定分割数据中准确的变点位置，同时实现变点个数与位置的精准估计。该方法既保留了 Jin 等提出的两阶段多变点估计方法“快速高效、精准定位”的优势，又通过指数平方损失函数抵御异常值干扰，为复杂数据环境下均值模型的多变点估计提供更可靠的解决方案。数值模拟和实证分析验证了该方法的有效性。

2. 基于指数平方损失函数的稳健均值多变点估计

2.1. 均值变点模型

考虑一个含有 s 个变点的均值模型，其中 $s\geq 2$，$1<a_1<\cdots<a_s<n$，模型如下

$$\boldsymbol{Y}=\begin{cases}\boldsymbol{\mu}_0+\varepsilon_i & 1\leq i\leq a_1,\\ \left(\boldsymbol{\mu}_0+\boldsymbol{\delta}_1\right)+\varepsilon_i & a_1<i\leq a_2,\\ \quad\vdots & \\ \left(\boldsymbol{\mu}_0+\sum_{l=1}^{s}\boldsymbol{\delta}_l\right)+\varepsilon_i & a_s<i\leq n.\end{cases} \tag{1}$$

其中 $\boldsymbol{Y}=\left(y_1,\cdots,y_n\right)^{\mathrm{T}}$ 是 n 维观测值，$\boldsymbol{\mu}_0$ 是初始值，s 是变点数量，$a_1,\cdots,a_s$ 是变点位置，$\boldsymbol{\delta}_l$ 是模型均值在变点处的增量，$\boldsymbol{\varepsilon}=\left(\varepsilon_1,\cdots,\varepsilon_n\right)^{\mathrm{T}}$ 是随机误差，$\boldsymbol{Y}$ 和 $\boldsymbol{\varepsilon}$ 均为 $n\times 1$ 维列向量。参考文献[13]，将数据 $\boldsymbol{Y}=\left(y_1,y_2,\cdots,y_n\right)^{\mathrm{T}}$ 分成 p_n+1 段，第一段长度为 $n-p_nm$，剩下 p_n 段长度为 m，段长 m 采用贝叶斯信息准则(Bayesian Information Criterion, BIC)选取，保证每个分段内至多有一个变点。

令 $\boldsymbol{X}$ 是一个下三角矩阵

$$\boldsymbol{X}=\left(\boldsymbol{I}^{(1)},\boldsymbol{I}^{(2)},\boldsymbol{I}^{(p_n+1)}\right)_{n\times(p_n+1)}=\begin{pmatrix}\boldsymbol{I}_{(1)} & 0 & 0 & \cdots & 0\\ \boldsymbol{I}_{(2)} & \boldsymbol{I}_{(2)} & 0 & \cdots & 0\\ \boldsymbol{I}_{(3)} & \boldsymbol{I}_{(3)} & \boldsymbol{I}_{(3)} & \cdots & 0\\ \vdots & \vdots & \vdots & \ddots & 0\\ \boldsymbol{I}_{(p_n+1)} & \boldsymbol{I}_{(p_n+1)} & \boldsymbol{I}_{(p_n+1)} & \cdots & \boldsymbol{I}_{(p_n+1)}\end{pmatrix}$$

其中 $\boldsymbol{I}_{(1)}$ 是 $n-p_nm$ 维单位矩阵，当 $j\in\{2,3,\cdots,p_n+1\}$ 时，$\boldsymbol{I}_{(j)}$ 是 m 维单位矩阵，则模型(1)可以转化为下面的矩阵形式

$$\boldsymbol{Y}=\boldsymbol{X}\boldsymbol{\theta}^{*}+\boldsymbol{X}_{\omega}\vec{\boldsymbol{\omega}}+\boldsymbol{\varepsilon} \tag{2}$$

关于 $\boldsymbol{X}_{\omega}$、$\vec{\boldsymbol{\omega}}$、$\boldsymbol{\theta}^{*}$ 的详细的定义以及模型(2)的推导详见文献[13]。这样，模型(1)中均值多变点的估计问题就转化为高维线性回归模型(2)的变量选择问题。

2.2. 指数平方损失函数

指数平方损失函数是一种具有稳健性的损失函数[23]，核心优点是它能够克服传统平方损失函数对异常值高敏感性的缺陷，同时保留了指数函数光滑性的优点。具体形式为

$$\varphi_\gamma = 1-\exp\left(-\left(y-\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}\right)^2/\gamma\right) \tag{3}$$

其中 $\gamma>0$ 是调节参数，用于控制估计量的稳健程度与估计效率。当 γ 较大时，由泰勒展开式可得 $\varphi_\gamma \approx \left(y-\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}\right)^2/\gamma$，此时所提出的估计量与最小二乘估计量相似；当 γ 较小时，损失函数趋近于常数 1，从而对参数的估计影响较小。因此，较小的 γ 会限制异常值对估计量的影响，降低估计量的敏感性。通常 γ 采用数据驱动的方法进行选择，从而在变量选择过程中同时获得高稳健性与高估计效率，详见文献[23]的 3.3 节。

由于 $\exp(\cdot)$ 函数是无限阶可导，梯度、Hessian 矩阵都很简洁，相比直接最小化 $1-\exp(\cdot)$，最大化 $\exp(\cdot)$ 函数的泰勒展开式、影响函数推导、渐近正态性的证明都更简单，且当残差越大的时候，$\exp(\cdot)$ 函数趋于 0 的速度更快，对异常值更稳健。因此 Wang (2013)对线性回归模型 $Y=\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}$ 提出基于指数平方损失函数的变量选择方法，目标函数为

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n}\exp\left\{-\left(Y_i-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)^2/\gamma\right\} - \sum_{j=1}^{d}P_\lambda\left(\left|\beta_j\right|\right) \tag{4}$$

通过最大化目标函数来求解参数估计。其中 d 是变量个数，$P_\lambda(\cdot)$ 是对变量系数的惩罚函数。惩罚函数中，$\lambda>0$ 为正则化参数，合适的 λ 可以将无关变量对应的系数 β_j 压缩至 0，从而实现变量选择。一般来说，λ 可采用多种方法选取，例如当惩罚函数 $P_\lambda\left(\left|\beta_j\right|\right)=\dfrac{\tau_j\left|\beta_j\right|}{\left|\tilde{\beta}_j\right|}$ 时，可通过最小化一个 BIC 型目标函数来选取最优 λ

$$\sum_{i=1}^{n}\left[1-\exp\left\{-\left(Y_i-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)^2/\gamma\right\}\right] + n\sum_{j=1}^{d}\frac{\tau_j\left|\beta_j\right|}{\left|\tilde{\beta}_j\right|} - \sum_{j=1}^{d}\log\left(0.5n\tau_j\right)\log(n) \tag{5}$$

其中 $\tilde{\beta}_j$ 是 β_j 的初始稳健估计，由此得到最优 $\lambda_j=\hat{\tau}_j/\left|\tilde{\beta}_j\right|$，这里 $\hat{\tau}_j=\log(n)/n$。Wang 等在文章中详细证明了该方法具有高崩溃点(近 1/2)，影响函数有界，且变量选择过程具有 Oracle 性质(模型选择的相合性、参数估计的渐近正态性)，具有良好的稳健性，详见文献[23]。

2.3. 基于指数平方损失的均值多变点的稳健估计

对模型(2)，文章采用指数平方损失函数加 L_1 惩罚，提出基于指数平方损失的两阶段多变点估计方法(TSMCD-ESL)，目标函数为

$$L\left(\boldsymbol{\theta}^*\right) = \sum_{i=1}^{n}\exp\left\{-\left(\boldsymbol{y}_i-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta}^*\right)^2/\gamma\right\} - \lambda\sum_{j=1}^{p_n+1}\sum_{i=1}^{q}\left|\theta_{ji}\right| \tag{6}$$

最大化目标函数 $L\left(\boldsymbol{\theta}^*\right)$ 得到 $\boldsymbol{\theta}^*$ 的相合估计量为

$$\hat{\boldsymbol{\theta}}^* = \arg\max_{\theta^*}\left\{\sum_{i=1}^{n}\exp\left\{-\left(\boldsymbol{y}_i-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta}^*\right)^2/\gamma\right\} - \lambda\sum_{j=1}^{p_n+1}\sum_{i=1}^{q}\left|\theta_{ji}\right|\right\} \tag{7}$$

令 $\boldsymbol{\theta}^*$ 的真实值为 $\boldsymbol{\theta}_0^* = \left(\boldsymbol{\theta}_{01}^{*\mathrm{T}}, \boldsymbol{\theta}_{02}^{*\mathrm{T}}\right)^{\mathrm{T}}$，其中 $\boldsymbol{\theta}_{01}^*$ 是真实非 0 系数，$\boldsymbol{\theta}_{02}^*$ 是真实 0 系数。(7)式的估计结果 $\hat{\boldsymbol{\theta}}^* = \left(\hat{\boldsymbol{\theta}}_1^{*\mathrm{T}}, \hat{\boldsymbol{\theta}}_2^{*\mathrm{T}}\right)^{\mathrm{T}}$，$\hat{\boldsymbol{\theta}}_1^*$ 是估计的非 0 系数，$\hat{\boldsymbol{\theta}}_2^*$ 是估计的 0 系数。由文献[23]可知，在适当的正则化条件下，$\hat{\boldsymbol{\theta}}^* = \left(\hat{\boldsymbol{\theta}}_1^{*\mathrm{T}}, \hat{\boldsymbol{\theta}}_2^{*\mathrm{T}}\right)^{\mathrm{T}}$ 满足 Oracle 性质，即

定理 1 当 $\Sigma = E\left(\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\right)$ 是正定的，且 $E\|\boldsymbol{x}\|^3<\infty$ 成立。有

(1) 相合性：$\left\|\hat{\boldsymbol{\theta}}^*-\boldsymbol{\theta}_0^*\right\| = O_p\left(n^{-1/2}\right)$

(2) 稀疏性：$\hat{\boldsymbol{\theta}}_2^* \xrightarrow{a.s.} 0$

(3) 渐近正态性：$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_1^*-\boldsymbol{\theta}_{01}^*\right) \xrightarrow{d} N(0,\Sigma)$

令 $\hat{\boldsymbol{\theta}}_1^*$ 对应的片段指标为集合 $\hat{\mathcal{A}}_n = \left\{j:\hat{\theta}_j\neq 0, j=1,\cdots,p_n+1\right\}$，

$\hat{\mathcal{A}}_n^* = \left\{ j : j \in \hat{\mathcal{A}}_n, j-1 \notin \hat{\mathcal{A}}_n, j = 2, \cdots, p_n + 1 \right\} = \left\{ \hat{k}_1, \cdots, \hat{k}_{\hat{t}} \right\}$，$\hat{t}$ 为初步估计的变点数量，$\hat{I}_{(l)} = I_{\hat{k}_l - 1} \bigcup I_{\hat{k}_l}$ 为第 l 个变点的候选分段。

为了准确估计出模型(1)的真实变点，采用文献[13]的第二阶段对候选片段进行精炼。令变点的真实位置为 a_l，根据上述分析，a_l 很可能位于候选分段 $\hat{I}_{(l)}$ 中，对于候选分段 $\hat{I}_{(l)} = \left\{ n_j^{(l)}, \cdots, n_j^{(r)} \right\} \left(l = 1, 2, \cdots, \hat{s} \right)$，采用拟似然比检验获取 a_l 的估计值 $\hat{a}_l$，具体步骤如下。

对每一个候选分段内构建含单个变点的线性回归模型

$$\boldsymbol{y}_i = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_1 I\left(n_j^{(l)} \le i \le \zeta_j \right) + \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_2 I\left(\zeta_j < i \le n_j^{(r)} \right) + \boldsymbol{\varepsilon}_i \tag{8}$$

其中 ζ_j 为可能的变点位置，$\boldsymbol{\beta}_1$ 和 $\boldsymbol{\beta}_2$ 为每一个预选分段内变点前后两段的回归系数。则 $\hat{I}_{(l)}$ 中是否存在变点等价于假设检验问题 $H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 \leftrightarrow H_1 : \boldsymbol{\beta}_1 \ne \boldsymbol{\beta}_2$。定义检验统计量：

$$T_l = N_l \left(\hat{\sigma}_l^2 - \min_{\beta_1} \sum_{i=n_j^{(l)}}^{\hat{\zeta}_j} \left(\boldsymbol{y}_i - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_1 \right)^2 - \min_{\beta_2} \sum_{i=\hat{\zeta}_j+1}^{n_j^{(r)}} \left(\boldsymbol{y}_i - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_2 \right)^2 \right) \Big/ \hat{\sigma}_j^2 \tag{9}$$

其中 N_l 为候选分段长度，$\hat{\sigma}_j^2 = \min_{\boldsymbol{\beta}} \sum_{i=n_j^{(l)}}^{n_j^{(r)}} \left(\boldsymbol{y}_i - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta} \right)^2$ 为无变点模型的残差平方和。检验统计量的临界值为：

$$b_l = \left(2 \ln \ln N_l + q \left(\ln \ln \ln N_l \right) / 2 - \ln \Gamma\left(\frac{q}{2} \right) \right)^2 \Big/ \left(2 \ln \ln N_l \right) \tag{10}$$

其中 $\Gamma(x) = \int_0^{+\infty} t^{x-1} \mathrm{e}^{-t} \mathrm{d}t \left(x > 0 \right)$ 是伽马函数，$c_l = \left(b_l / \left(2 \ln \ln N_l \right) \right)^{1/2}$，根据 Csörgő & Horváth (1997)的极限定理可得，若 $T_l > b_l + 2c_l \log\left(-2 / \log\left(1 - \alpha \right) \right)$ 时，则认为分段内存在变点，其中 α 为置信水平，取值为 0.05。则变点位置的估计量为：

$$\hat{\zeta}_l = \arg \min_{n_j^{(l)} + q < k < n_j^{(r)} - q} \left[\min_{\beta_1} \sum_{i=n_j^{(l)}}^{k} \left(\boldsymbol{y}_i - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_1 \right)^2 + \min_{\beta_2} \sum_{i=k+1}^{n_j^{(r)}} \left(\boldsymbol{y}_i - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_2 \right)^2 \right] \tag{11}$$

经过假设检验已获得变点的粗略候选集 $\left\{ \hat{\zeta}_1, \hat{\zeta}_2, \cdots, \hat{\zeta}_{\hat{k}} \right\}$，然而初始均匀分段可能导致变点位置存在局部偏移，且含变点的分段拟合会引入估计偏差。为提升变点检测的位置精度与一致性进行第二次精细筛选，令 $\left\{ \hat{\tau}_1, \hat{\tau}_2, \cdots, \hat{\tau}_{\hat{k}} \right\}$ 为粗略候选集中变点在全局样本的位置，用集合 $\left\{ 0, \hat{\tau}_1, \hat{\tau}_2, \cdots, \hat{\tau}_{\hat{k}}, n \right\}$ 中的数据将样本重新分成 $k+1$ 段，此时每个新的子段内部为无变点的平稳序列，子段间的边界即为变点的潜在位置。在每两个相邻子段上进行第二次拟似然比检验，剔除假阳性变点并优化位置估计，最终得到变点估计集合 $\mathcal{A}_n = \left\{ \hat{a}_1, \cdots, \hat{a}_{\hat{s}} \right\}$。由 Csörgő & Horváth (1997)的极限定理可知，对于分段平稳数据，变点的粗略估计具有 $O_p(1)$ 的收敛率，而在局部窗口内的拟似然比检验筛选的变点可实现变点位置的“超收敛”。

2.4. 变点估计的相合性

为了研究变点数量及其位置估计量的性质：

假设 1： 如果 $s \ge 1$，对于 $1 \le j \le s$，$a_{j,n} / n \to \tau_j > 0$，进一步地，当 $s \ge 2$ 时，有 $\min_{1 \le j \le s-1} \left(\tau_{j+1} - \tau_j \right) > 0$，保证变点间距足够大。

假设 2： 当 $t_2 - t_1 \to \infty$ 时，$\frac{1}{t_2 - t_1} \sum_{i=t_1}^{t_2} \boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}} \to W$，其中 W 是一个严格正定矩阵。

假设 3： 误差项满足：$\{ \varepsilon_i \}$ 为独立同分布随机变量序列，$E\left(\varepsilon_i \right) = 0$，$Var\left(\varepsilon_i \right) = \sigma^2 < \infty$，且 $E\left(\left| \varepsilon_i \right|^4 \right) < \infty$。

假设 4： 指数平方损失函数的光滑性：$E\left(\exp\left(-\frac{\epsilon_i^2}{\gamma} \right) \boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}} \right)$ 为有限矩阵，且 $E\left(\exp\left(-\frac{\epsilon_i^2}{\gamma} \right) \left(\frac{2\epsilon_i^2}{\gamma} - 1 \right) \boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}} \right)$ 为负定矩阵。

定理 2：在满足假设 1~假设 4 的情况下，$\hat{s}$ 为 TSMCD-ESL 方法估计的变点数量，s 为真实变点数量，$\hat{a}_l\left(l=1,\cdots,\hat{s}\right)$ 为 TSMCD-ESL 方法估计的变点位置，$a_l\left(l=1,\cdots,s\right)$ 为真实变点位置，则有：

(1) $\lim\limits_{n\to\infty}P\left(\hat{s}=s\right)=1$

(2) $\lim\limits_{n\to\infty}P\left(a_l\in\hat{I}_{(l)},\forall l=1,2,\cdots,s\,|\,\hat{s}=s\right)=1$

注：定理 2 的结论(1)说明变点估计的数量是相合的，结论(2)说明当估计的变点数量相合时，真实变点 a_l 位于 $\hat{I}_{(l)}$ 的概率为 1。定理 2 的详细证明参考文献[13]及其补充材料。

Wang 等在文献[23]中详细证明利用指数平方损失函数进行变量选择方法是具有 Oracle 性质的，TSMCD-ESL 方法的第一阶段将均值变点模型(1)转化为变量选择模型(2)，利用指数平方损失函数进行变量选择得到的估计量 $\hat{\boldsymbol{\theta}}^*$ 自然具有 Oracle 性质。结合定理 2，说明 TSMCD-ESL 方法所得到的变点估计是具有相合性和一致性的。

3. 数值模拟

运用数值模拟的方法，对样本量 $n=100$ 和 $n=1000$ 分别设置了无异常值、含有 1%异常值、5%异常值、10%异常值四种情况，有异常值数据中 y_i 基础值种子设置为 123，异常值位置随机产生。从估计精度和运算时间两方面对比了 TSMCD-ESL 方法与未分段的累积和方法，以及 Jin 2016 提出的基于平方损失函数两阶段多变点估计方法 $\text{TSMCD}_{(\text{adaptive lasso})}$，并重复 1000 次实验。累积和方法使用 R 中自带的“changepoint”包实现，Jin 2016 提出的基于平方损失函数两阶段多变点估计方法使用 R 中自带的“TSMCP”包实现。

3.1. 小样本情况下的多变点稳健估计

考虑初始数据集是一个含有 2 个变点的均值变点模型，样本量 n 为 100，真实变点位置为 $1<a_1=30<a_2=75<100$，即均值回归模型形式如下：

$$y_i=\begin{cases}\boldsymbol{\mu}_0+\varepsilon_i & 1\le i<30,\\ \left(\boldsymbol{\mu}_0+\boldsymbol{\delta}_1\right)+\varepsilon_i & 30\le i<75,\\ \left(\boldsymbol{\mu}_0+\boldsymbol{\delta}_1+\boldsymbol{\delta}_2\right)+\varepsilon_i & 75\le i\le 100.\end{cases}\tag{12}$$

其中，$\boldsymbol{\mu}_0=4$，$\boldsymbol{\delta}_1=-5$，$\boldsymbol{\delta}_2=3$。设 $\{\varepsilon_i\}$ 服从标准正态分布 $N\left(0,0.5\right)$。在这个数据集中分别随机替换 1%、5%、10%的异常值，异常值服从正态分布 $N\left(0,15\right)$。参考文献[13]数值模拟过程中每个子段的长度 $m_l=\kappa_l\sqrt{n},(l=1,2,\cdots,L)$，其中 $\text{TSMCD}_{(\text{adaptive lasso})}$ 中 κ_l 取值为 0.3，TSMCD-ESL 中 κ_l 取值为 0.7。

图 1 给出模型(12)在无异常值、含有 1%异常值、5%异常值、10%异常值四种情况下使用 TSMCD-ESL 方法的一次估计结果。红色圆点表示随机加入的异常值，蓝色虚线为估计变点的位置，红色虚线为真实变点的位置。

表 1 给出了模型(12)在无异常值、含有 1%异常值、5%异常值、10%异常值四种情况下 1000 次独立重复试验得到的变点位置的估计结果，其中 $\left\{\left|\hat{a}_i-a_i\right|\le 5\right\}$ 表示三种方法得到的估计变点位置与真实变点位置差距小于等于 5 的次数，“运行时间”表示三种方法 1000 次独立重复试验运行时间的均值。

从图 1 和表 1 可以看出，TSMCD-ESL 方法展现出良好的稳健性，在所有异常值比例下，均保持较高的检测准确率，显著优于其他两种方法。尽管其运行时间略逊于另外两种方法，但在实际含噪声数据场景中，这种效率牺牲换来了可靠的变点识别能力。

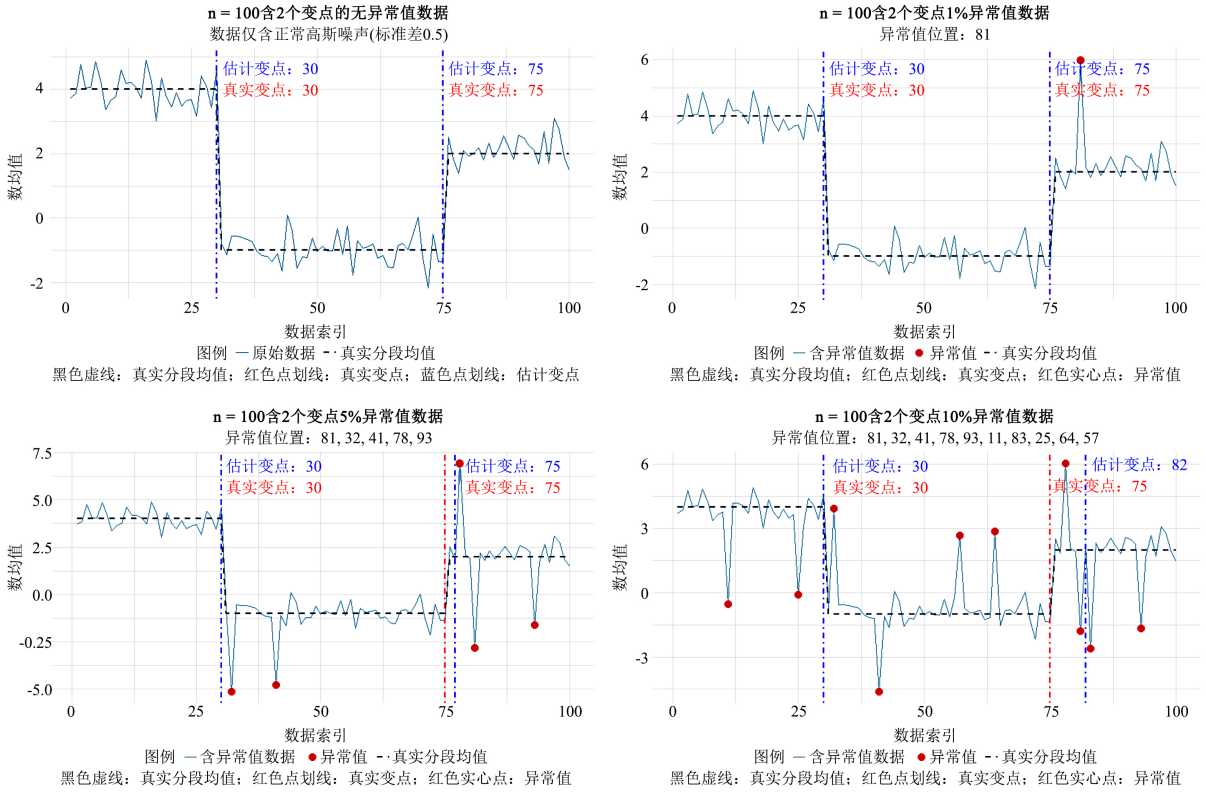

Figure 1. Results of a single simulation using the TSMCD-ESL method with $n = 100$ outliers under four different outlier scenarios
图 1. $n = 100$ TSMCD-ESL 方法 4 种异常值情况下模拟一次的结果

Table 1. Numerical simulation results for $n = 100$
表 1. $n = 100$ 数值模拟结果

方法	$\{\lvert\hat{a}_1 - 30\rvert \le 5\}$	$\{\lvert\hat{a}_2 - 75\rvert \le 5\}$	运行时间
无异常值			
累积和方法	1000	1000	0.0003
TSMCD$_{(\text{adaptive lasso})}$方法	981	523	0.0233
TSMCD-ESL 方法	987	970	0.0272
1%异常值			
累积和方法	733	296	0.0003
TSMCD$_{(\text{adaptive lasso})}$方法	894	393	0.0243
TSMCD-ESL 方法	1000	952	0.0273
5%异常值			
累积和方法	219	2	0.0004
TSMCD$_{(\text{adaptive lasso})}$方法	595	136	0.0202
TSMCD-ESL 方法	900	702	0.0266

续表

10%异常值			
累积和方法	45	0	0.0003
TSMCD$_{(\text{adaptive lasso})}$方法	325	34	0.0169
TSMCD-ESL 方法	748	458	0.0276

3.2. 大样本情况下的变点估计

考虑一个含有 3 个变点的均值回归模型，样本量 n 为 1000，真实变点位置为 $1<a_1=330<a_2=530<a_3=810<1000$，即均值回归模型形式如下：

$$y_i=\begin{cases}\boldsymbol{\mu}_0+\varepsilon_i & 1\le i<330,\\ \left(\boldsymbol{\mu}_0+\boldsymbol{\delta}_1\right)+\varepsilon_i & 331\le i<500,\\ \left(\boldsymbol{\mu}_0+\boldsymbol{\delta}_1+\boldsymbol{\delta}_2\right)+\varepsilon_i & 501\le i<810,\\ \left(\boldsymbol{\mu}_0+\boldsymbol{\delta}_1+\boldsymbol{\delta}_2+\boldsymbol{\delta}_3\right)+\varepsilon_i & 811\le i\le 1000.\end{cases} \tag{13}$$

其中 $\boldsymbol{\mu}_0=4$，$\boldsymbol{\delta}_1=-5$，$\boldsymbol{\delta}_2=3$，$\boldsymbol{\delta}_3=-4$。设$\{\varepsilon_i\}$服从标准正态分布 $N(0,0.5)$。在这个数据集中分别随机替换 1%、5%、10%的异常值，异常值服从正态分布 $N(0,15)$。数值模拟过程中每个子段的长度 $m_l=\kappa_l\sqrt{n},(l=1,2,\cdots,L)$，其中 TSMCD$_{(\text{adaptive lasso})}$中 κ_l 取值为 0.3。TSMCD-ESL 中 κ_l 取值为 0.7。

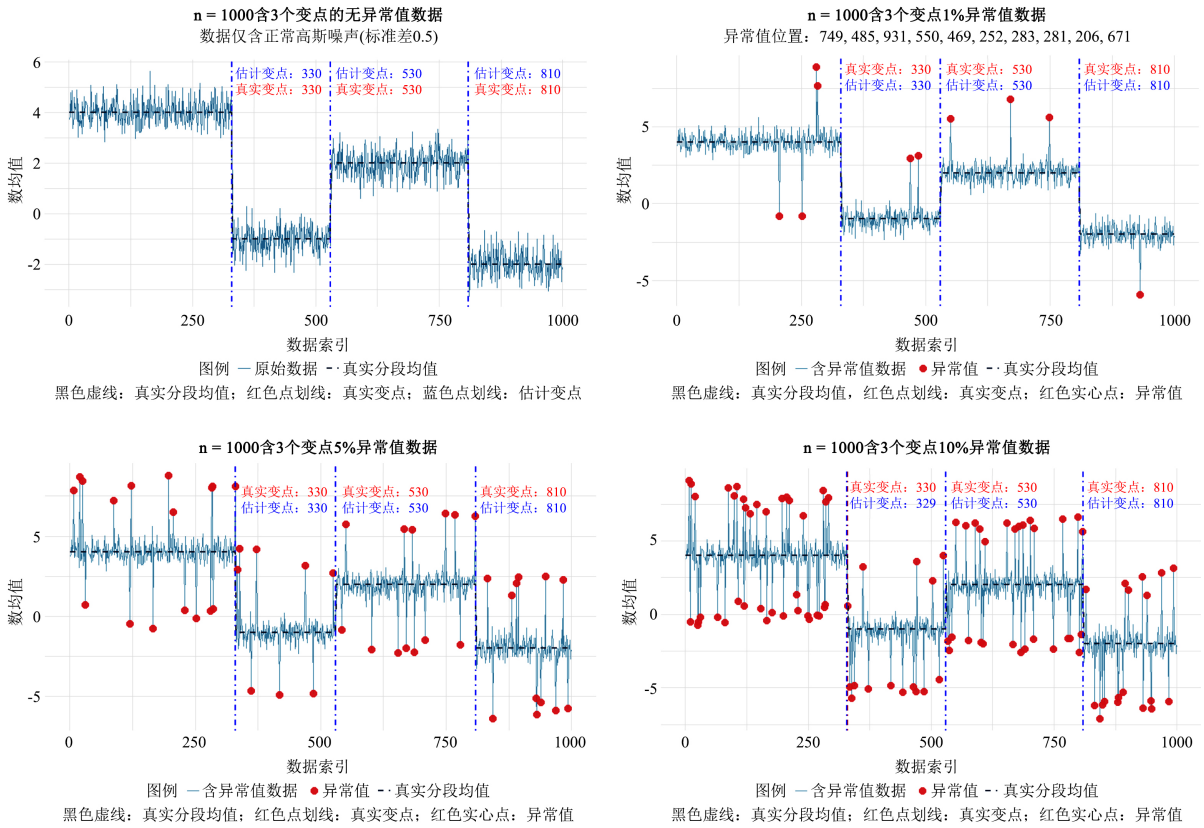

Figure 2. Results of a single simulation using the TSMCD-ESL method with n = 1000 under four outlier conditions
图 2. n = 1000 TSMCD-ESL 方法 4 种异常值情况下模拟一次的结果

图 2 给出模型(13)在无异常值、含有 1%异常值、5%异常值、10%异常值四种情况下使用 TSMCD-ESL 方法的一次估计结果。红色圆点表示随机加入的异常值，蓝色虚线为估计变点的位置，红色虚线为真实变点的位置。

表 2 给出了模型(13)在无异常值、含有 1%异常值、5%异常值、10%异常值四种情况下 1000 次独立重复试验得到的变点位置的估计结果，其中 $\{|\hat{a}_i - a_i| \le 5\}$ 表示三种方法得到的估计变点位置与真实变点位置差距小于等于 5 的次数，“运行时间”表示三种方法 1000 次独立重复试验运行时间的均值。

Table 2. Numerical simulation results for $n = 1000$
表 2. $n = 1000$ 数值模拟结果

方法	$\{\lvert\hat{a}_1 - 330\rvert \le 5\}$	$\{\lvert\hat{a}_2 - 530\rvert \le 5\}$	$\{\lvert\hat{a}_3 - 810\rvert \le 5\}$	运行时间
无异常值				
累积和方法无	1000	1000	1000	0.0045
TSMCD$_{(\text{adaptive lasso})}$方法	1000	1000	998	0.1723
TSMCD-ESL 方法	979	960	981	0.0694
1%异常值				
累积和方法	23	0	0	0.0036
TSMCD$_{(\text{adaptive lasso})}$方法	987	758	760	0.1724
TSMCD-ESL 方法	942	845	923	0.0746
5%异常值				
累积和方法	0	0	0	0.0056
TSMCD$_{(\text{adaptive lasso})}$方法	883	195	205	0.1652
TSMCD-ESL 方法	668	470	678	0.0767
10%异常值				
累积和方法	0	0	0	0.0082
TSMCD$_{(\text{adaptive lasso})}$方法	624	40	30	0.1341
TSMCD-ESL 方法	426	257	465	0.0135

综合图 1、图 2 和表 1、表 2 可以看出，无论是小样本还是大样本场景下，在无异常值的理想环境中，累积和方法凭借简单的计算逻辑，实现了变点 100%的检测准确率与最低的运行时间，是无噪场景下的高效方案。但面对实际数据中常见的异常值干扰，其估计性能较低。在含有 1%的异常值情况下基本丧失多变点估计能力，含异常值 5%及以上比例时无法识别任何变点，实用性局限极强。TSMCD$_{(\text{adaptive lasso})}$方法在低比例异常值下能维持一定精度，但随异常值增加，对后序变点的估计能力急剧衰退，难以应对高干扰场景。相比之下，TSMCD-ESL 方法展现出更均衡的性能，在各类异常值比例下，均能保持三种方法中最高的综合估计准确率，尤其在 5%、10%高比例异常值下，对变点的识别稳定性显著优于累计和方法和 TSMCD$_{(\text{adaptive lasso})}$方法。

4. 实证分析

选取 R 包“changepoint。influence”中自带的测井数据集进行实证分析，该数据集包含 4050 组地下岩石的核磁响应测量值，这些数据是通过将探测仪下入钻孔，在离散时间点采集得到的，其记录的信息可反映岩石的结构特征，尤其是岩层分层情况。Ruanaidh [24]以及 Fearnhead [25]的早期研究中，均采用均值变点模型进行分析，分别检测出 13 个和 16 个变点。文章应用 TSMCD$_{(\text{adaptive lasso})}$和 TSMCD-ESL 两种方法对该数据集进行变点估计，并对结果进行对比，结果见表 3 和图 3、图 4。

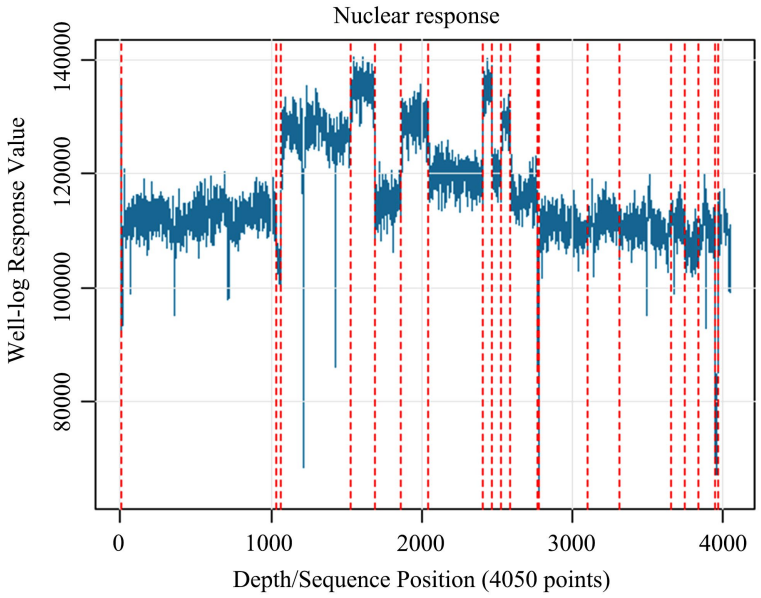

Figure 3. Change point estimation results using the TSMCD$_{(\text{adaptive lasso})}$ method
图 3. TSMCD$_{(\text{adaptive lasso})}$方法变点估计结果

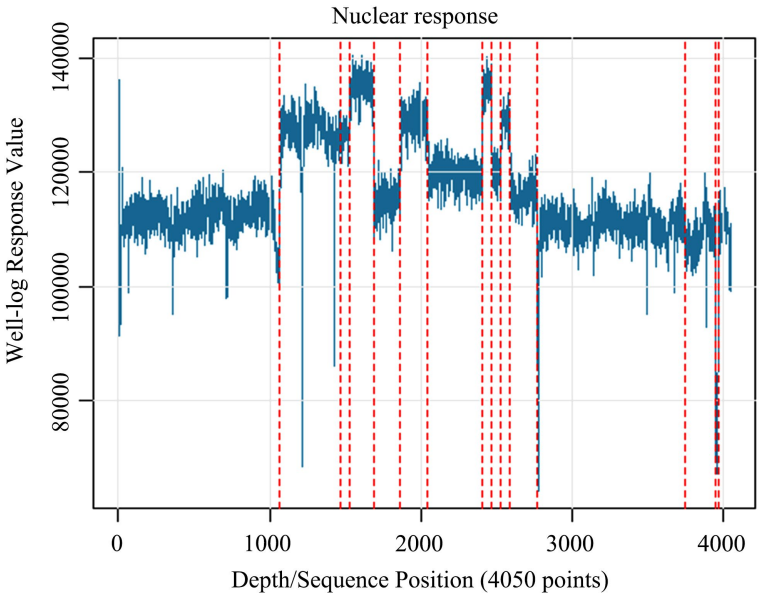

Figure 4. Change point estimation results of the TSMCD-ESL method
图 4. TSMCD-ESL 方法变点估计结果

Table 3. Estimation results of change points in well logging data using two methods
表 3. 两种方法测井数据变点估计结果

方法	变点个数	变点位置						
$TSMCD_{(adaptive\ lasso)}$	20	6	1034	1070	1526	1685	1866	2047
		2409	2469	2531	2591	2768	2779	3103
		3314	3656	3744	3841	3944	3963	
TSMCD-ESL	14	1070	1465	1526	1685	1866	2046	2408
		2469	2530	2591	2772	3744	3942	3965

由表 3 和图 3、图 4 可以看出 $TSMCD_{(adaptive\ lasso)}$方法共估计出 20 个变点，涵盖 6、1034、1070、1526 等位置；TSMCD-ESL 方法估计出 14 个变点，具体位置包括 1070、1465、1526 等。两种方法在核心变点识别上具有较高一致性，1070、1526、1685、1866、2469、2591、3744 等关键变点均被精准捕捉，印证了这些点位是测井数据中真实存在的均值突变位置。但是 $TSMCD_{(adaptive\ lasso)}$方法估计出变点数量更多，额外识别出 6、1034、3314、3656 等变点位置，由图 3 可以看出这几个位置极有可能是由异常值引起的虚假变点，整体数据结构并没有改变；TSMCD-ESL 方法则通过指数平方损失的约束作用，过滤了部分弱信号变点，估计结果更为稳健，在兼顾估计精度的同时，有效降低了虚假变点的检出概率。

5. 结论

文章聚焦均值模型的多变点估计问题，创新性地提出了融合指数平方损失函数的两阶段多变点检测方法(TSMCD-ESL)。该方法通过数据分段与变量选择锁定潜在变点区域，结合拟似然比检验实现变点精准定位，借助指数平方损失函数的特性有效抵御异常值干扰，兼顾了检测效率与稳健性。

未来研究可从三方面进一步拓展：一是拓展模型的适用场景，将 TSMCD-ESL 方法推广至广义线性模型、面板数据模型等更复杂的统计模型，适配金融、环境等领域的多元数据变点估计需求；二是优化参数选择策略，针对不同数据特征设计自适应调参算法，无需人工预设子段长度等参数，提升方法的通用性；三是继续深化理论与应用研究，完善高维数据、相依数据场景下的变点估计相合性证明，同时开展更多跨领域实证分析，如地震监测、基因序列分析等，进一步验证方法的实用价值，为复杂数据的结构性突变分析提供更高效的统计工具。

基金项目

新疆维吾尔自治区自然科学基金项目(2023D01A37)。

参考文献

[1] 李美琪, 金百锁, 董翠玲. 线性回归模型中相依数据的多结构变点的估计[J]. 中国科学: 数学, 2023, 53(7): 1007-1024.

[2] Pepelyshev, A. and Polunchenko, A.S. (2017) Real-Time Financial Surveillance via Quickest Change-Point Detection Methods. *Statistics and Its Interface*, **10**, 93-106. https://doi.org/10.4310/sii.2017.v10.n1.a9

[3] Chen, J. and Gupta, A.K. (2014) Parametric Statistical Change Point Analysis: With Applications to Genetics, Medicine, and Finance. Birkhäuser.

[4] Ghosh, P. and Vaida, F. (2010) Random Changepoint Modelling of HIV Immunologic Responses. *Statistics in Medicine*, **26**, 2074-2087. https://doi.org/10.1002/sim.2671

[5] Goncalves, A.M. (2013) Change-Point Analysis in Environmental Time Series. Repositório Institucional da Universidade de

Aveiro, 1-12.

[6] Kim, J. and Cheon, S. (2009) Multiple Change-Point Estimation of Air Pollution Mean Vectors. *Korean Journal of Applied Statistics*, **22**, 687-695. https://doi.org/10.5351/kjas.2009.22.4.687

[7] PAGE, E.S. (1954) Continuous Inspection Schemes. *Biometrika*, **41**, 100-115. https://doi.org/10.1093/biomet/41.1-2.100

[8] Hinkley, D.V. (1971) Inference about the Change-Point from Cumulative Sum Tests. *Biometrika*, **58**, 509-523. https://doi.org/10.1093/biomet/58.3.509

[9] Csörgö, M. and Horváth, L. (1997) Limit Theorems in Change-Point Analysis. Wiley.

[10] Wu, Y. (2005) Inference for Change Point and Post Change Means after a CUSUM Test. Springer.

[11] Brodsky, B.E. and Darkhovsky, B.S. (1993) Nonparametric Methods in Change-Point Problems. Kluwer Academic Publishers.

[12] Harchaoui, Z. and Lévy-Leduc, C. (2010) Multiple Change-Point Estimation with a Total Variation Penalty. *Journal of the American Statistical Association*, **105**, 1480-1493. https://doi.org/10.1198/jasa.2010.tm09181

[13] Jin, B., Wu, Y. and Shi, X. (2016) Consistent Two-Stage Multiple Change-Point Detection in Linear Models. *Canadian Journal of Statistics*, **44**, 161-179. https://doi.org/10.1002/cjs.11282

[14] 邹航, 姜云卢. 高维线性回归模型稳健变量选择方法综述[J]. 应用概率统计, 2024, 40(1): 157-181.

[15] Huber, P.J. (1981) Robust Statistics. Wiley. https://doi.org/10.1002/0471725250

[16] Yohai, V.J. (1987) High Breakdown-Point and High Efficiency Robust Estimates for Regression. *The Annals of Statistics*, **15**, 642-656. https://doi.org/10.1214/aos/1176350366

[17] Yohai, V.J. and Zamar, R.H. (1988) High Breakdown-Point Estimates of Regression by Means of the Minimization of an Efficient Scale. *Journal of the American Statistical Association*, **83**, 406-413. https://doi.org/10.1080/01621459.1988.10478611

[18] Cuzick, J. (1988) Rank Regression. *The Annals of Statistics*, **16**, 1369-1389. https://doi.org/10.1214/aos/1176351044

[19] Rousseeuw, P.J. (1984) Least Median of Squares Regression. *Journal of the American Statistical Association*, **79**, 871-880. https://doi.org/10.1080/01621459.1984.10477105

[20] Rousseeuw, P.J. and Leroy, A.M. (1987) Robust Regression and Outlier Detection. Wiley. https://doi.org/10.1002/0471725382

[21] Rousseeuw, P. and Yohai, V. (1984) Robust Regression by Means of S-Estimators. In: Franke, J., Härdle, W. and Martin, D., Eds., *Lecture Notes in Statistics*, Springer US, 256-272. https://doi.org/10.1007/978-1-4615-7821-5_15

[22] Fan, J. and Li, R. (2001) Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *Journal of the American Statistical Association*, **96**, 1348-1360. https://doi.org/10.1198/016214501753382273

[23] Wang, X., Jiang, Y., Huang, M. and Zhang, H. (2013) Robust Variable Selection with Exponential Squared Loss. *Journal of the American Statistical Association*, 108, 632-643. https://doi.org/10.1080/01621459.2013.766613

[24] Ruanaidh, J.J.K. and Fitzgerald, W.J. (1996) Numerical Bayesian Methods Applied to Signal Processing. Springer.

[25] Fearnhead, P. and Clifford, P. (2003) On-Line Inference for Hidden Markov Models via Particle Filters. *Journal of the Royal Statistical Society Series B*: *Statistical Methodology*, **65**, 887-899. https://doi.org/10.1111/1467-9868.00421