第48卷第11期 中国科学技术大学学报 Vol. 48, No. 11
2018年11月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Nov. 2018

文章编号:0253-2778(2018)11-0869-08

基于稳健S估计的长江流域气象异常值检测

金百锁,李炽坤

(中国科学技术大学管理学院统计与金融系,安徽合肥 230026)

摘要:高维数据如气象数据中不可避免地存在异常值,应用最广泛的最小二乘法在识别异常值上不具有稳健性和灵敏度.稳健估计方法可使求出的估计量不受异常数据的强烈影响,从而能更好地识别异常点.这里给出了基于稳健S估计的主成分分析模型,其中加入Tukey的双权型函数约束条件.该模型无须对数据分布函数的具体形式做假设,算法的收敛速度较快.之后再结合B样条函数对数据作平滑处理,以平均残差平方和为检验统计量,使用同样具有稳健性的调优箱型图作为判别异常值的界限.实证分析采用了我国长江流域5个城市60多年共约58 000条气象数据,分别运用PCA方法和基于稳健S估计的异常值判别方法对该数据集进行了对比分析.可以明显地看出,相比传统方法,基于稳健S估计的异常值判别方法更突出地给出关于异常值的信息,能更好地识别异常值.

关键词:稳健估计;主成分分析;异常值检测;高维数据;降维

中图分类号:C812 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2018.11.001

2010 Mathematics Subject Classification:Primary 62G35;Secondary 62G05

引用格式:金百锁,李炽坤. 基于稳健S估计的长江流域气象异常值检测[J]. 中国科学技术大学学报,2018,48(11):869-876.

JIN Baisuo, LI Chikun. Outlier detection of Yangtze River basin meteorological data based on robust S-estimator[J]. Journal of University of Science and Technology of China, 2018,48(11):869-876.

Outlier detection of Yangtze River basin meteorological data based on robust S-estimator

JIN Baisuo, LI Chikun

(*Department of Statistics and Finance, School of Management, University of Science and Technology of China, Hefei* 230026, *China*)

Abstract: Outlier is unavoidable in high-dimensional data, such as meteorological data, and the the most widely used least-square method has no robustness and sensitivity in detecting outliers. Robust estimation can make the estimators not strongly influenced by outliers, so that the outliers can be better identified. By adding Tukey's biweight function constraints, a principal component analysis model based on robust S-estimator was established, which converges rapidly and does not need to assume the specific form of the distribution function. Then the observations were smoothed by B-spline basis, the mean residuals squared norm was used as the test statistic, and the adjusted box-plot which also has robustness was trained to detect the outliers. In the example, more than 58 thousand measurements of meteorological data over 60 years of 5 cities in Yangtze River basin were adopted. A comparative analysis of the data set with outlier

收稿日期:2018-04-18;**修回日期**:2018-06-15

基金项目:国家自然科学基金(11571337)资助.

作者简介:金百锁(通讯作者),男,1980年生,博士/副教授.研究方向:变点,空间统计. E-mail:jbs@ustc.edu.cn

detecting procedure based on principal component analysis and robust S-estimator has been conducted. It can be seen clearly that compared with the classical approach, the outlier detecting procedure based on robust S-estimator gives more information on the abnormal data, and thus can identify outliers better.

Key words: robust estimation; principal component analysis; outlier detection; high-dimensional data; dimension reduction

0 引言

极端气候事件,如局部地区短时间内的持续高温酷热,发生旱灾或者洪涝等,可以视为温度湿度降雨量等气象数据发生异常. 21世纪以来,在全球变暖的背景下,短期气候异常愈加频繁,众多破纪录气候事件给社会带来了严重的影响,因而对极端气候事件的演变特征、发生机理及影响评估已成为近年来全球研究热点. 况雪源等[1]提出了一个基于气象学理论的识别群发性极端气候事件的方法,并应用于我国近50年来群发性高温事件的识别,对相应事件的频率分布、年代际差异及区域特征作了总结分析. 对这些异常现象的研究可以为理解气候灾害的变化规律及预测评估提供客观的参考依据,亦可为短期气候预测及防灾减灾工作提供理论参考. 有鉴于此,本文尝试从另外的角度,通过统计方法构建异常点检测模型来研究我国气象数据的异常问题,另一方面,也为新的异常值检测方法提供实证依据.

气象数据具有时间上的连续性、空间上的相关性以及波动小分布较为集中等特征,要对其做异常检测,首先需要以固定时间段为维度,收集不同地区的大量数据进行分析估计寻找出数据的主要特征. 主成分分析法(PCA(principal component analysis)方法)在处理这类问题上具有一定的优势,主要以均值向量、协方差或相关系数矩阵为基础,压缩冗余信息,降低数据维度,从而得到合适的估计值. 但由于使用的统计量的不稳健性,当数据包含少数异常值时,由PCA方法得到的估计值很容易产生较大的偏差. 较早的研究主要从两个方面去改善PCA方法,一是使用稳健散布矩阵,如MCD(minimum covariance determinant)[2],稳健S估计(robust S-estimator)[3],代替原来的协方差矩阵;二是利用高维数据的投影追踪(projection-pursuit)[4]方法,增加一个约束条件:在降维的同时使目标函数(projection index)最大化. 近期的研究如Hubert等[5]提出一种基于稳健估计的主成分分析法,综合了稳健散布矩阵和投影追踪法的特点,使数据集未污染时得到的估计量更为精确,而对污染数据的估计更为稳健,并将新方法用于化学和工程的数据异常诊断;Bali等[6]主要从理论上将投影追踪法拓展到函数型数据的PCA方法中,完善了一些证明并做了模拟测试.

考虑到稳健统计方法拟合的残差可以更少偏倚,更突出地给出关于异常值的信息,能更好地识别异常值,我们将在PCA方法的框架下,使用稳健S估计代替协方差矩阵,加入带损失函数的约束条件,通过迭代法求解获得观测数据的稳健估计;同时引入B样条做原始数据到函数型数据的转换,而后基于稳健估计的残差项构建检验统计量识别异常气象数据,并与PCA方法做对比分析.

1 模型及算法

作为处理复杂高维数据的一种有效方法,主成分分析法将原始数据映射到方差贡献率最高的几个维度上,设法将原来众多具有一定相关性的指标,重新组合成一组新的互相无关的综合指标,要求它们尽可能多地保留原始数据的信息且彼此互不相关,这些综合指标也就是主成分. 而后通过主成分来揭示原数据集的内部结构,达到降维和简化问题的目的. 本文的模型借鉴主成分分析和最小二乘法(OLS(ordinary least square)方法)的策略,以稳健估计方法为切入点,尝试构建残差项的稳健估计量,目的仍是使估计结果的残差平方和最小.

记 $\boldsymbol{x}_i$, $i=1,\cdots,n$,为原始 p 维观测数据,共同组成 $n\times p$ 维设计矩阵 $\boldsymbol{X}$,$\hat{\boldsymbol{x}}_i$ 是 $\boldsymbol{x}_i$ 的估计量(文中向量均为列向量);$\|\cdot\|$ 代表欧几里德范数,$\boldsymbol{A}=(\boldsymbol{a}_1,\boldsymbol{a}_2,\cdots,\boldsymbol{a}_n)^{\mathrm{T}}$ 是 $n\times q$ 维因子载荷矩阵,$\boldsymbol{a}_i$ 为它的行向量;$\boldsymbol{B}=(b_1,b_2,\cdots,b_p)^{\mathrm{T}}$ 是 $p\times q$ 维基矩阵($q<p$),$\boldsymbol{b}_j$,$j=1,\cdots,p$,为它的行向量,且 $\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B}=\boldsymbol{I}$,即各组基之间相互正交;$\mu$ (location estimator)是 p 维向量,包含了观测数据的位置信息;估计值 $\hat{\boldsymbol{x}}_i=\boldsymbol{B}\boldsymbol{a}_i+\boldsymbol{\mu}$,$\boldsymbol{r}_{ij}=\boldsymbol{x}_{ij}-\hat{\boldsymbol{x}}_{ij}$ 为残差项.

$$L_o(\boldsymbol{A},\boldsymbol{B},\boldsymbol{\mu})=\sum_{i=1}^{n}\|\boldsymbol{x}_i-\hat{\boldsymbol{x}}_i\|^2=\sum_{i=1}^{n}\sum_{j=1}^{p}\boldsymbol{r}_{ij}{}^2 \tag{1}$$

式(1)是求解 OLS 估计的初始公式，按照 OLS 方法的思想，最优的估计值应满足

$$\hat{\boldsymbol{x}}_i^o = \underset{\hat{\boldsymbol{x}}_i \in \mathbb{R}^p}{\operatorname{argmin}} L_o(\boldsymbol{A},\boldsymbol{B},\boldsymbol{\mu}) \tag{2}$$

现记 $\boldsymbol{X}$ 的协方差矩阵为 $\boldsymbol{\Sigma}$，就可以围绕 $\boldsymbol{\Sigma}$ 进行主成分的相关计算. 例如可以通过奇异值分解，解出 $\boldsymbol{\Sigma}=\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\mathrm{T}}$，其中 $\boldsymbol{U},\boldsymbol{D},\boldsymbol{V}$ 均为 $p\times p$ 的方阵，$\boldsymbol{U},\boldsymbol{V}$ 为正交矩阵，$\boldsymbol{D}$ 为 $\boldsymbol{\Sigma}$ 的奇异值组成的对角阵. 取左奇异矩阵的前 q 列作为基矩阵 $\boldsymbol{B}$，$\boldsymbol{\mu}$ 取均值向量 $\boldsymbol{\mu}=n^{-1}\sum_{i=1}^{n}\boldsymbol{x}_i$，即得使式(2) 成立的 $\boldsymbol{a}_i^{\mathrm{T}}=(\boldsymbol{x}_i-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{B}$.

为了减少观测数据中的异常值对残差项 r_{ij} 的影响，引入 r_{ij} 的稳健估计 $\hat{\sigma}_j$，常用下面式(3)来求解，通过该式得出的 $\hat{\sigma}_j$ 也称为 r_{ij} 的一个稳健 S 估计(robust scale estimator)[7].

$$\frac{1}{n}\sum_{i=1}^{n}\rho_c\left(\frac{r_{ij}(\boldsymbol{A},\boldsymbol{B},\boldsymbol{\mu})}{\hat{\sigma}_j}\right)=b \tag{3}$$

式中，损失函数 ρ 一般要求有界可导，这里选定 $\rho(y)=\min(3y^2-3y^4+y^6,1)$ (Tukey 双权函数). c 和 b 都是调节常数(turning constant)，$\rho_c(\mu)=\rho(\mu/c)$，当残差项 r 服从特定分布 $\boldsymbol{F}(\mu_1,\sigma^2)$ 时，c 用来保证 $\hat{\sigma}_j$ 收敛到 σ，而 $b=E(\rho((r-\mu_1)/\sigma))$. 这里选择 $c=3.0$，$b=0.2426$，以保证残差项服从正态分布时 $\hat{\sigma}_j$ 的 Fisher 一致性[8]. 假定残差项 r_{ij} 已知，可以通过牛顿迭代法求出式(3)中的 $\hat{\sigma}_j$：

①取初始值 $\hat{\sigma}_j^0=\underset{i}{\mathrm{MAD}}(r_{ij})+1$，MAD 为绝对中位差函数，常作为分布尺度(scale)参数的估计量. 这样的初值选择法可以在保证 $\hat{\sigma}_j^0$ 稳健性的同时使算法较容易收敛；

②计算 $\hat{\sigma}_j^1=\hat{\sigma}_j^0(bn)^{-1}\sum_{i=1}^{n}\rho_c(r_{ij}/\hat{\sigma}_j^0)$；

③用 $\hat{\sigma}_j^1$ 替代步骤②等式右边的 $\hat{\sigma}_j^0$，得到 $\hat{\sigma}_j^2=\hat{\sigma}_j^1(bn)^{-1}\sum_{i=1}^{n}\rho_c(r_{ij}/\hat{\sigma}_j^1)\cdots$ 重复直至 $|\hat{\sigma}_j^{k+1}-\hat{\sigma}_j^k|<\epsilon$ 时停止迭代，ϵ 为一足够小的常数(本文 $\epsilon=1\times10^{-6}$). 取 $\hat{\sigma}_j=\hat{\sigma}_j^{k+1}$ 是为解.

得到残差项的稳健估计后，将其代入式(1)右边，接下来问题变为求解

$$\hat{\boldsymbol{x}}_i^s = \underset{\hat{\boldsymbol{x}}_i \in \mathbb{R}^p}{\operatorname{argmin}} L_s(\boldsymbol{A},\boldsymbol{B},\boldsymbol{\mu}) \tag{4}$$

式中，

$$L_s(\boldsymbol{A},\boldsymbol{B},\boldsymbol{\mu})=\sum_{j=1}^{p}\hat{\sigma}_j^2 \tag{5}$$

仿照 PCA 方法的结果，同样可以得到估计值 $\hat{\boldsymbol{x}}_i^s$，不同的是这里基矩阵 $\boldsymbol{B}$ 和位置向量 $\boldsymbol{\mu}$ 均未知. 考虑一般最值问题的解法，将 $\boldsymbol{r}_{ij}(\boldsymbol{A},\boldsymbol{B},\boldsymbol{\mu})=\boldsymbol{x}_{ij}-\hat{\boldsymbol{x}}_{ij}^s=\boldsymbol{x}_{ij}-\boldsymbol{\mu}_j-\boldsymbol{a}_i^{\mathrm{T}}\boldsymbol{b}_j$ 代入式(3)，并把常数 c 放入损失函数 ρ 中，尝试对向量 $\boldsymbol{a}_i$ 求导：

$$\frac{\partial}{\partial \boldsymbol{a}_i}\left(\sum_{j=1}^{p}\hat{\sigma}_j^2\right)=\sum_{j=1}^{p}2\hat{\sigma}_j\frac{\partial\hat{\sigma}_j}{\partial a_i}=-2\sum_{j=1}^{p}h_j^{-1}\rho'\left(\frac{r_{ij}}{\hat{\sigma}_j}\right)b_j \tag{6}$$

式中，$h_j=\hat{\sigma}_j^{-1}\sum_{i=1}^{n}\rho'(r_{ij}/\hat{\sigma}_j)r_{ij}/\hat{\sigma}_j$. 同样的分别再对 $\boldsymbol{b}_j,\boldsymbol{u}_j$ 求导，令这三个偏导数都等于 0，整理可得

$$\sum_{j=1}^{p}\boldsymbol{w}_{ij}(\boldsymbol{x}_{ij}-\boldsymbol{u}_j)\boldsymbol{b}_j=\left(\sum_{j=1}^{p}\boldsymbol{w}_{ij}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathrm{T}}\right)\boldsymbol{a}_i \tag{7}$$

$$\sum_{i=1}^{n}\boldsymbol{w}_{ij}(\boldsymbol{x}_{ij}-\boldsymbol{u}_j)\boldsymbol{a}_i=\left(\sum_{i=1}^{n}\boldsymbol{w}_{ij}\boldsymbol{a}_i\boldsymbol{a}_i^{\mathrm{T}}\right)\boldsymbol{b}_j \tag{8}$$

$$\sum_{i=1}^{n}\boldsymbol{w}_{ij}(\boldsymbol{x}_{ij}-\boldsymbol{a}_i^{\mathrm{T}}\boldsymbol{b}_j)=\sum_{i=1}^{n}\boldsymbol{w}_{ij}\boldsymbol{u}_j \tag{9}$$

式中，权重系数 $\boldsymbol{w}_{ij}=(\boldsymbol{h}_j\boldsymbol{r}_{ij})^{-1}\boldsymbol{\rho}'(\boldsymbol{r}_{ij}/\hat{\boldsymbol{\sigma}}_j)$.

求解式(4)以及式(7)～(9)，是一个较为复杂的非凸优化问题，普通方法较为困难. 参考文献[8]提出的迭代加权最小二乘法(iterative re-weighted least squares，IRWLS)，将求解过程分为以下 4 步：

①从 $\boldsymbol{X}^{\mathrm{T}}$ 中随机选出 q 列组成新的设计矩阵 $\boldsymbol{X}^{(1)}$ (相当于 PCA 方法中的协方差矩阵 $\boldsymbol{\Sigma}$)，将 $\boldsymbol{X}^{(1)}$ 通过 $\boldsymbol{QR}$ 分解变为正交矩阵 $\boldsymbol{Q}$ 与上三角矩阵 $\boldsymbol{R}$ 的乘积，取初始基矩阵 $\boldsymbol{B}=\boldsymbol{Q}$. 参照上文使用迭代法求解式(3)中 $\hat{\sigma}_j$ 时第 1 步代入值 $\hat{\sigma}_j^0$ 的选择方法，为保证稳健性，初始的 μ 取 $\underset{\mu}{\operatorname{argmin}}\sum_{i=1}^{n}\|x_i-\mu\|$，由 $a_i^{\mathrm{T}}=(x_i-\mu)^{\mathrm{T}}\boldsymbol{B}$，以及 $\hat{x}_i^s=\boldsymbol{B}a_i+\mu$，将 $\mu,\boldsymbol{B}$ 代入，得到 a_i，初始的估计 $x_i^{(0)}$ 以及残差 $r_{ij}^{(0)}$；

②把步骤①中得到的 $r_{ij}^{(0)}$ 代入式(3)，解出对应的 S 估计 $\hat{\sigma}_j^{(0)}$，再将 $r_{ij}^{(0)},\hat{\sigma}_j^{(0)}$ 代入 ρ'，算出 $w_{ij}^{(0)}$；将 $w_{ij}^{(0)}$，步骤①中基矩阵 $\boldsymbol{B}$ 的行向量 b_j 以及 x_{ij},μ_j，一起代入式(7)，算出 $a_i^{(0)}$；同样将 $a_i^{(0)}$ 等代入式(8)，算出 $b_j^{(0)}$，从而得到新的基矩阵 $\boldsymbol{B}^{(0)}$；再将 $a_i^{(0)},b_j^{(0)}$ 等代入式(9)，算出 $\mu_j^{(0)}$；

③由步骤②中得到的 $a_i^{(0)},\boldsymbol{B}^{(0)},\mu_j^{(0)}$，可以算出新的估计 $\hat{x}_i^{(1)}$ 以及残差 $r_{ij}^{(1)}$，代入式(3)，解出对应的 S 估计 $\hat{\sigma}_j^{(1)}\cdots$；重复以上各步，直到

$$\left|\sum_{j=1}^{p}(\hat{\sigma}_j^{(k+1)})^2-\sum_{j=1}^{p}(\hat{\sigma}_j^{(k)})^2\right|<\epsilon$$

时停止迭代;取 $\hat{\sigma}_j=\hat{\sigma}_j^{(k+1)}$,返回对应的估计值 $\hat{x}_i^s=\boldsymbol{B}^{(k)}a_i^{(k)}+\mu^{(k)}$;

④这样每次从 $\boldsymbol{X}^{\mathrm{T}}$ 中随机选出 q 列组成新的设计矩阵 $\boldsymbol{X}^{(\cdot)}$,都可以得出一组相应残差的稳健估计 $\hat{\sigma}_j$. 重复选取 N 次,综合每组得到的 $\sum_{j=1}^{p}\hat{\sigma}_j^2$,取其值最小的那一组为最终的结果.

显然,N 的值不应超过 $\binom{n}{q}$,而 q 的值在 PCA 方法中一般依靠主成分的方差贡献率来确定. 实际问题中,两者的选择还需要兼顾算法复杂度和结果的可靠性. 在一定量的测试和比较不同的取值后,发现对于测试集选取 $N=200$,$q=2$,最大迭代次数为 500 次,可以在减少计算时间的同时得到满意的结果.

由于求解上述迭代问题的过程中需要不断对特定矩阵求逆,而且直接降维很可能会造成求解结果失真,故考虑对原数据集做 B 样条近似. 这样将降维过程分为两步,更不容易丢失原数据的主要特征,且平滑化后的数据代入上述迭代算式更容易在一定迭代次数内收敛.

B 样条函数的主要特点是局部存在紧致支集,以至于近似的函数型数据在局部具有非常好的平滑性. 更重要的是,只要较少的节点(knots),就可以对函数型数据提供良好的近似,且有较高的计算效率和稳定性[9]. 由于立方样条函数已有较好的近似结果,为此,本文采用四阶均匀 B 样条函数. 设节点个数为 k($k<p$),求出样条基的转换矩阵 $\boldsymbol{K}$($p\times k$ 维),易得系数向量 $\boldsymbol{y}_i^{\mathrm{T}}=\boldsymbol{x}_i{}^{\mathrm{T}}\boldsymbol{K}$,以 $\boldsymbol{y}_i$ 作为新的样本数据,通过前述方法求出 $\boldsymbol{y}_i$ 的估计值 $\hat{y}_i$,然后返回 $\hat{\boldsymbol{x}}_i=\boldsymbol{K}\hat{y}_i$. 实证分析中将均匀地取连续区间 $[0,1]$ 内的 $k=20$ 个节点,使用 R 语言 mgcv 程序包中 cSplineDes 函数构建 B 样条基. 为方便计算,使迭代过程更容易收敛,同样对 $\boldsymbol{K}$ 作正交分解 $\boldsymbol{K}=\boldsymbol{QR}$,最后取 $\boldsymbol{Q}$ 作为样条基的转换矩阵 $\boldsymbol{K}$,即 $\boldsymbol{y}_i^{\mathrm{T}}=\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{Q}$.

2 异常值检测

我们将给出异常值的判别方法,包括判别界限的确定,检验统计量,以及异常值的判别流程. 首先,箱型图作为描述性统计分析和数据可视化最基础的工具之一,直观地表示了数据的位置、散布、偏态、厚尾等信息,可以作为高维数据异常值的检测方法. 箱型图的上下须(whisker)给出了区间:

$$[Q_1-1.5\mathrm{IQR},Q_3+1.5\mathrm{IQR}] \tag{10}$$

式中,Q_1,Q_3 分别表示数据的 25%,75%分位点(下同),$\mathrm{IQR}=Q_3-Q_1$ 称为四分位间距,式(10)以外的点便可以划为异常点. 这样用是否超出箱型图的界限来判别异常值,是可行和简便的. 但是,当测试数据服从偏态分布时,就会有很多超出箱型图上下须的"正常数据"被误判为异常值. 因此需要在式(10)中加入数据的偏态等信息. 文献[10]针对性地提出了调优的箱型图法(adjusted boxplot)推荐区间

$$[U_1,U_2]=\begin{cases}[Q_1-c\mathrm{e}^{-bM}\mathrm{IQR},Q_3+c\mathrm{e}^{-aM}\mathrm{IQR}], & M<0;\\ [Q_1-c\mathrm{e}^{aM}\mathrm{IQR},Q_3+c\mathrm{e}^{bM}\mathrm{IQR}], & M\geqslant 0\end{cases} \tag{11}$$

替代式(10)的上下界. 式中,a,b,c 为调节常数,常设置 $a=-4$,$b=3$,$c=1.5$(详见 R 程序 robustbase 程序包中 adjbox 函数). 而

$$M=\underset{x_i\leqslant Q_2\leqslant x_j}{\mathrm{median}}h(x_i,x_j)$$

是衡量分布偏度的估计量,可以看作是三阶矩的一个替代,不同的是它的计算基于样本中位数,因而更具稳健性. Q_2 是样本中位数,h 是核函数,一般可以取

$$h(x_i,x_j)=\frac{(x_j-Q_2)-(Q_2-x_i)}{x_j-x_i},x_i\neq x_j.$$

显然,原数据 x_i 偏离我们得到的估计值 $\hat{x}_i$ 越多,越有可能是异常值. 现计算平均残差平方和,构建检验统计量

$$\gamma_i=\frac{1}{j}\sum_{j=1}^{p}(\hat{\boldsymbol{x}}_{ij}-x_{ij})^2 \tag{12}$$

γ_i 的分布是未知的,估计起来也比较复杂,通过式(11)来判别异常值,无须对 γ_i 的分布做假定. 除调优的箱型图法外,还有其他一些较为方便的判别法,如文献[11]针对函数型数据提出的 HDR(high-density region),BAG(functional bagplots)方法. 文献[8]就 R 语言 rainbow 程序包中的六种不同的异常值判别法(包括 HDR,BAG)与调优的箱型图法做了比较模拟研究,结果也倾向于后者. 据此,由 γ_i 的基本性质,并加入对 x_i 的区间估计,最终的判别过程分为两步:

①由节 1 的估计法得出 $\hat{x}_i$,代入式(12),计算各 γ_i 和式(11)中的判别区间. 考虑 γ_i 的正则属性,若 $\gamma_i>U_2$,则判定 x_i 异常,其余情况都记作正常. 应用中仅凭此步骤易发生漏判,为继续找出潜在的

异常点，添加了下一步；

②若所有 x_i 都标为正常，可以假定数据集服从正态分布，容易列出它在第 j 维均值的 $1-\alpha$（如取 $\alpha=0.05$）置信区间为 $[a_j, d_j]$. 其中，

$$d_j = l_j + s_j/\sqrt{p}\cdot t_{(1-\alpha/2)}(p-1),$$

$$l_j = 1/n\cdot\sum_{i=1}^{n} x_{ij},$$

$$s_j = \sqrt{1/(n-1)\cdot\sum_{i=1}^{n}(x_{ij}-l_j)^2},$$

$t_{(1-\alpha/2)}(p-1)$ 为自由度为 $p-1$ 的 t 分布上 $1-\alpha/2$ 分位点，$\#\{\cdot\}$ 表示集合中元素个数，$I\{\cdot\}$ 为示性函数. 得到截断的平均残差平方和

$$\gamma_i^0 = \frac{1}{\#\{j, x_{ij}\geqslant d_j\}}\cdot\sum_j(\hat{x}_{ij}-x_{ij})^2\boldsymbol{I}\{x_{ij}\geqslant d_j\} \tag{13}$$

用 γ_i^0 代替 γ_i 代入步骤①，返回结果.

3　实证分析

实证分析对象数据来自国家气象信息中心（中国气象局气象数据中心）[12]. 我们希望通过运用基于稳健 S 估计的主成分分析法和 PCA 方法做识别异常值的对比分析，发现标的城市的夏季气象异常变化，验证新方法在处理此类问题上的优良性质. 数据来源是我国长江流域五个主要城市的地面观测站，即安庆（58424），重庆（沙坪坝 57516），南京（58238），上海（宝山 58362），武汉（57494），小括号中数字对应的是各地面观测站点编号. 测试数据集由这五个城市 1954～2016 年的 6～8 月共计 92 天每日的最高气温，以及 1951～2016 年的 6～8 月每日的相对湿度组成（对应图 1 和图 2(a)中的各观测点，其中上海是 1959～2016 年的数据，包括气温和湿度，不再单独说明）. 原始数据集基于国家地面基础气象资料建设项目归档的“1951～2016 年中国国家级地面站数据更正后的月报数据文件（A0/A1/A）基础资料集”研制而成，制作过程经过严格质量控制，各要素的质量及完整性相对于以往发布的地面同类数据明显提高，正确率均接近 100%，各要素项数据的实有率普遍在 99%以上，对于缺失数据利用近邻 5 日的中位数估计.

我们使用的数据集包含了连续 63 年 5796 天的温度和 66 年 6072 天的湿度数据，均保留两位小数，统计结果如表 1 所示. 表 1 中，每日平均相对湿度数值如 0.45 代表湿度 45%；最后两列为污染数据，用测试 1，2 表示，分别为武汉 2016 年 6～8 月的每日平均气温和最小相对湿度，对应各行数据均显著低于武汉总体. 不难看出，各地区的气温分布较为接近，最高气温在 40℃左右，单日最高气温为 43℃（重庆），平均在 31.5℃上下，重庆和武汉的 75%分位点分别为 35.6℃和 34.6℃，说明高温天气相对较多；各城市温度均值普遍低于中位数 0.4℃左右，说明其分布是左偏的；而湿度的这两项统计指标都很接近 0.78，最小值在重庆（0.29），最大值均接近 1，相比温度，各城市湿度更为接近，分布更加平稳，某一年度发生异常的可能性更小. 相关性分析指出，单个城市温度和湿度之间存在强的负相关关系，以重庆为例，两要素间相关系数为 −0.82，显著性检验 P 值小于 1×10^{-15}；不同城市单要素两两之间，则大多呈高度正相关，以安庆和南京为例，相关系数分别为 0.86（温度），0.70（湿度），检验显著.

表 1　五个城市气象数据统计结果

Tab. 1　Summary statistics of meteorological data in five cities

	每日最高气温/(℃)					每日平均相对湿度					污染数据	
	安庆	重庆	南京	上海	武汉	安庆	重庆	南京	上海	武汉	测试 1	测试 2
最小值	15.9	18.60	18.30	19.00	19.30	0.45	0.29	0.38	0.47	0.37	19.00	0.28
25%分位点	29.00	29.20	28.60	27.30	29.60	0.72	0.68	0.74	0.76	0.72	25.00	0.50
中位数	31.80	32.70	31.50	30.50	32.40	0.78	0.77	0.80	0.82	0.78	27.60	0.56
均值	31.48	32.23	31.11	30.13	31.92	0.78	0.76	0.79	0.81	0.78	27.74	0.60
75%分位点	34.30	35.60	33.80	33.00	34.60	0.85	0.84	0.86	0.87	0.85	30.95	0.68
最大值	40.90	43.00	40.70	39.90	39.60	1.00	1.00	0.99	0.98	1.00	33.00	0.97

对各个城市的每日最高气温分别使用新方法和 PCA 方法估计. 以武汉气温数据为例，取夏季的每日最高气温为 x_{ij}，得到 $n\times p$ 设计矩阵 $\boldsymbol{X}$，其中 $n=63$ 为总年数，维数 $p=92$ 为夏季总天数，通过

样条基变换,将日数据转换为对应每一年份的函数型数据,计算得到用于诊断的残差统计量 γ_i,将结果记录在表2中,若通过 γ_i 检测 x_i 为异常,则将该年的气温标为异常.为测试基于稳健S估计的异常值判别方法和PCA方法的差别,在不影响数据集整体特性的情况下,单独对武汉2016年使用污染数据(表1测试1列).其中S和PCA两列记录了诊断统计量 γ_i 的值.而异常这一列,数字0表示该年为正常年份,1代表仅用S估计量判别为异常,2代表用两种方法都判别为异常,由于篇幅限制,对于所有城市温度都正常的年份不做记录.图1是五个城市的温度曲线,仅用S估计量判别为异常的年份被单独标出.气象业务上将单站日最高气温超过35℃的天气称为高温天气,而25℃则为对于人体相对最为舒适的推荐温度,也是夏季偏低的温度,故将这两个数值在图1中用虚线标出.

表2 温度数据的异常值检测

Tab.2 Outlier detection of temperature data

年份	安庆			重庆			南京			上海			武汉		
	S	PCA	异常	S	PCA	异常	S	PCA	异常	S	PCA	异常	S	PCA	异常
1959	5.16	5.13	0	8.74	9.17	0	13.02	7.94	1	7.34	7.47	0	9.19	8.82	0
1961	4.91	5.02	0	18.02	13.88	1	4.04	4.36	0	4.37	4.44	0	5.16	5.03	0
1966	7.25	6.30	0	7.46	11.47	0	16.02	9.89	1	7.01	7.24	0	9.33	9.46	0
1967	5.38	6.02	0	6.60	7.85	0	14.52	9.90	1	6.45	6.34	0	8.63	7.90	0
1980	17.52	11.58	1	15.72	13.26	0	13.34	10.72	1	15.08	11.98	2	17.19	12.44	1
1992	7.92	8.56	0	22.90	19.62	2	10.36	7.56	0	6.35	7.04	0	7.11	8.46	0
1993	14.38	10.68	1	9.93	9.34	0	9.04	8.97	0	11.65	9.79	0	8.91	9.46	0
2002	16.11	13.97	2	11.35	10.00	0	11.59	10.60	0	10.35	9.75	0	12.69	10.70	0
2003	10.28	10.75	0	15.21	14.33	0	11.49	13.04	0	9.64	10.16	0	14.39	14.08	2
2005	11.6	11.75	0	8.64	10.22	0	12.26	14.11	2	9.37	10.31	0	16.48	13.94	2
2008	8.81	8.47	0	18.37	12.38	1	5.78	8.47	0	6.67	6.63	0	7.30	7.25	0
2009	12.94	11.89	0	16.98	14.51	0	13.22	12.18	2	11.95	11.22	0	9.92	10.24	0
2011	10.65	12.46	0	5.37	12.17	0	12.77	12.75	2	9.68	9.82	0	6.51	7.38	0
2013	10.28	9.20	0	20.76	14.05	1	8.56	8.62	0	12.82	8.98	1	7.74	7.02	0
2014	11.04	8.63	0	19.61	17.45	1	9.62	9.36	0	6.95	6.30	0	9.83	9.85	0
2016	8.64	8.24	0	15.74	14.79	0	6.75	7.16	0	7.08	5.78	0	13.09	8.53	1

况雪源等[1]研究了我国近50年来群发性高温事件的统计特征,识别出1961～2012年群发性高温事件510次,列出了综合强度位于前10的高温事件,其中包括1966年7月16日至1966年8月17日,1967年7月14日至1967年8月13日,2011年7月20日至2011年8月9日,并且有5次发生在2000年以后,占了50%.表2新方法识别的异常年份中,包含了以上三个高温事件发生的时间段,且2000年以后的占比略超50%,符合较好,总体也与夏季高温热浪增强和全球变暖的趋势相一致.

具体对于图1:安庆1980年,1993年低于 Q_1 的天数占比高达47.8%,40.2%,其中1980年为所有年份最高;重庆2014年低温天数偏多,占比17.39%为所有年份最高,1961年,2008年,2013年的高温天数比例为55.43%,32.61%,60.87%,其中2013年为所有年份最高;南京1959年,1966年,1967年的高温天数比例分别为27.17%,40.22%,32.61%,1966年为所有年份最高,7月中下旬气温逼近40℃高温,原因是该年份长江流域南京段夏季发生了历史罕见的高温大旱(详见地方县志记载),而1980年虽然高温天数较少(7.61%),但温度低于 Q_1 的天数占比48.91%,仅次于1954年的52.17%;对于上海,除去1980年的"凉夏",新方法还将2013年判为异常,其高温天数占比47.83%,为所有年份最高;最后是武汉,新方法多识别的异常年份为1980年和2016年.2016年低温天数占比26.09%为所有年份最高,且其夏季最高温出现在7月31日(33℃),没有超过35℃,原因是该年份我们使用的是武汉日平均气温,实际其7月31日最高气温达到38.4℃,PCA方法没有检测出这个异常,表中对应的 γ_i 值8.53属于正常范围,S估计量给出的值高达13.09.以上各城市对于两种方法都检测为异常

的年份不再多做说明，例如武汉 2003 年，2005 年的高温天数比例分别为 35.87%，34.78%.

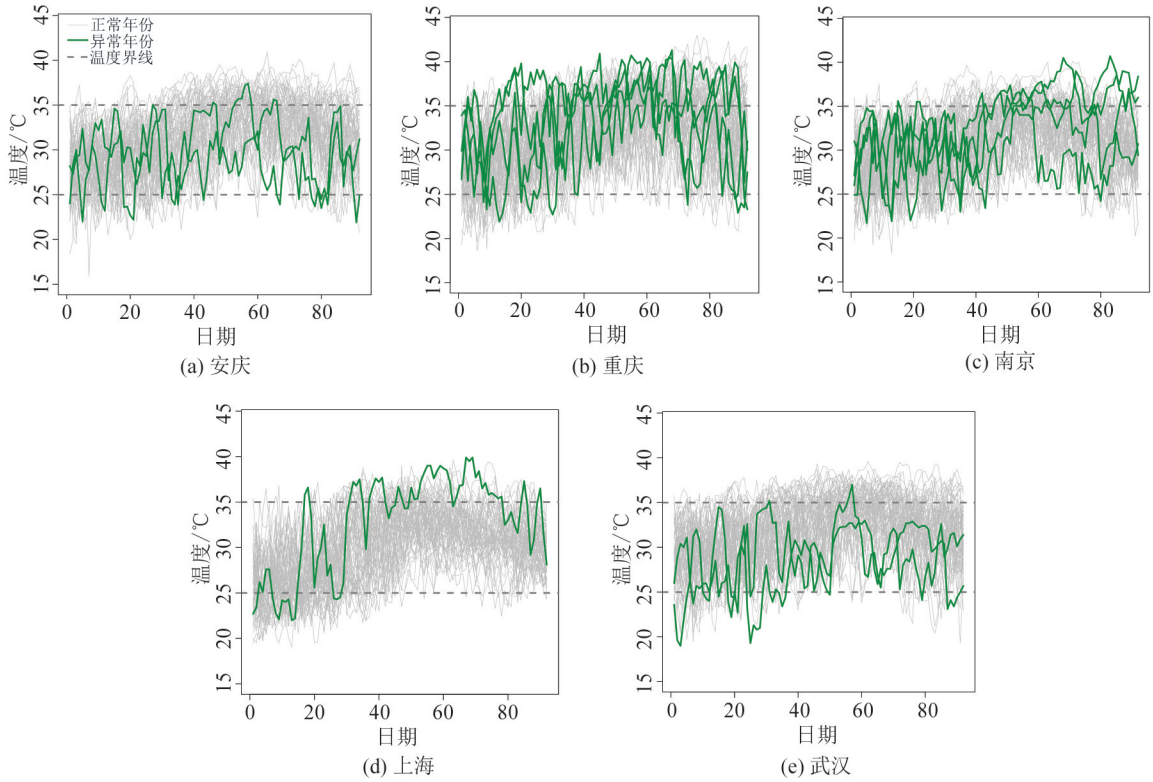

图 1　各城市夏季气温及异常年份

Fig. 1　Summer temperature and abnormal years in five cities

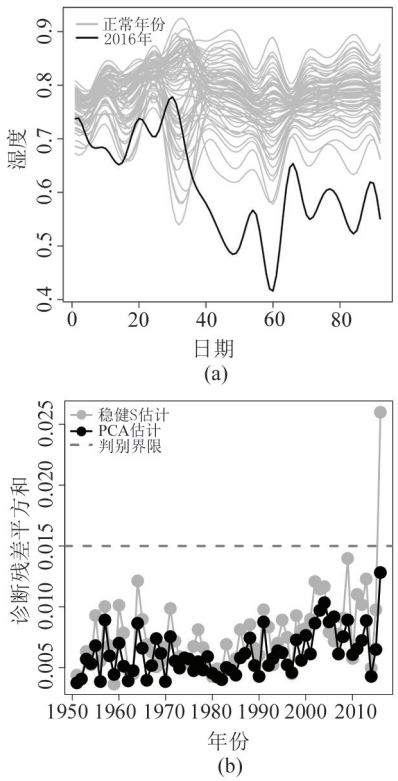

图 2　武汉湿度检测

Fig. 2　Detection of humidity data for Wuhan

接下来检测每个城市的每日平均相对湿度. 同样为测试两种方法的差别，单独对武汉 2016 年使用日最低相对湿度(表 1 测试 2 列). 安庆，仅由 S 估计检测出 2012 年为异常，诊断统计量为 15.91，该年度 23.91%的天数湿度偏低(低于 60%)，为测试所有年份中最高；重庆，两种方法都检测出的年份有 1961 年，1992 年，2013 年，仅由新方法检出的有 2008 年，2011 年，实际 2011 年，2013 年湿度偏低的天数占比为 44.57%，57.61%，超过其他年份，而 2008 年湿度普遍偏高，高于 90%的有 17 天；南京，仅由新方法判别出 2013 年为异常，该城市湿度偏低的天数平均有 3.44%，而此年度达到 20.65%. 上海，两种方法均没有发现异常值；最后是武汉，图 2(a)作出了由新方法得到的 $\hat{\boldsymbol{x}}_i^S$，污染数据离群较为明显，由于使用的是最低相对湿度，2016 年度 61.96%的天数湿度低于 60%，为测试所有年份中最高. PCA 方法同样没能识别出异常年份，图 2(b)列出了所有年份的诊断残差平方和，该年度的诊断统计量分别为 0.0265，0.0128，基于新方法得到的 γ_i 值明显偏高.

综上，新方法检测到了所有 PCA 方法找到的异

常年份,且对后者未能识别的污染数据检验显著,就其余其单独检出的异常年份而言,与实际符合较好,且在一定程度上概括了原始数据的相关性特征:如1980年有4个城市检测出异常,原因是夏季整体温度偏低;温度诊断为异常的年份,湿度往往也显示异常,如重庆温度异常的5个年份中,有4个湿度也检验为异常.湿度总体检出的异常年份相对较少,这点也与数据整体的平稳特征相符.

4 结论

本文在主成分分析模型的框架下,将OLS估计需要最小化的残差项替代为相应的稳健S估计量,采用加权迭代法求解带约束的最小二乘问题,使得到的估计值在满足既定损失函数的同时残差平方和也最小.在不损失主要信息的情况下提取出数据的主要特征,构建合适的检验统计量,引入加入偏态信息的判别区间,从而构成完整的基于稳健估计的异常值诊断模型,对自20世纪50年代起连续60余年我国长江流域五个城市夏季的两种气象数据进行了检测,并与基于PCA方法的异常值诊断结果进行了比较.研究得到的结论如下:

(Ⅰ)理论上,新方法构建的稳健估计量相比普通估计量有更高的破坏点,判别阈值也考虑了数据的偏态信息,且新方法的判别过程采用了两步法,因而能消除多个异常值的遮蔽作用,使得异常值能够正确地被识别出来;

(Ⅱ)实际数据的分析中,新方法在诊断效率和准确性上都优于PCA方法,不仅能识别出后者未发现的污染数据,而且能在表现相对平稳的气象数据中检测出其他异常年份,检测结果解释力良好,能与实际和相关研究文献相符合.对群发性高温事件的识别、灾害天气的诊断和预防能提供一定的有价值的参考;

(Ⅲ)新方法无须对原数据的分布做假定,不仅可以用于气象数据的异常值检测上,还可以推广到其他类似的复杂高维数据的分析中,例如可以应用到制造业产品质量检测,金融市场风险监控,信用卡欺诈检测,医疗图像分析处理和网络入侵检测等其他研究领域;虽然文中求解式(4)及式(5)得到的稳健S估计量具有优良的相合性和渐进正态性,但是式(3)的构造和式(5)的解法并不唯一,使用其他方法能否能够给出稳健性更佳的估计量以用于异常值判别,都有待后续进一步的研究.

参考文献(References)

[1] 况雪源,王遵娅,张耀存,等. 中国近50年来群发性高温事件的识别及统计特征[J]. 地球物理学报,2014,57(6):1782-1791.
KUANG Xueyuan, WANG Zunya, ZHANG Yaocun, et al. Identification and statistical characteristics of the cluster high temperature events during last fifty years [J]. Chinese Journal of Geophysics, 2014, 57(6): 1782-1791.

[2] ROUSSEEUW P. Least median of squares regression [J]. Journal of the American Statistical Association, 1984, 79(388): 871-880.

[3] ROUSSEEUW P, YOHAI V. Robust Regression by Means of S-Estimators [M]. New York: Springer, 1984: 256-272.

[4] LI G, CHEN Z. Projection-pursuit approach to robust dispersion matrices and principal components: Primary theory and Monte Carlo[J]. Journal of the American Statistical Association, 1985, 80(391): 759-766.

[5] HUBERT M, ROUSSEEUW P J, BRANDEN K V. ROBPCA: A new approach to robust principal component analysis[J]. Technometrics, 2005, 47(1): 64-79.

[6] BALI J L, BOENTE G, TYLER D E, et al. Robust functional principal components: A projection-pursuit approach[J]. The Annals of Statistics, 2011, 39(6): 2852-2882.

[7] MARONNA R, MATIN D, YOHAI V. Robust Statistics: Theory and Methods [M]. Chichester, England: John Wiley, 2006: 1-84.

[8] BOENTE G, SALIBIAN M. S-estimators for functional principal component analysis[J]. Journal of the American Statistical Association, 2015, 110(511): 1100-1111.

[9] HE X, ZHU Z, FUNG W K. Estimation in a semiparametric model for longitudinal data with unspecified dependence structure [J]. Biometrika, 2002, 89(3): 579-590.

[10] HUBERT M, VANDERVIEREN E. An adjusted boxplot for skewed distributions [J]. Computational Statistics & Data Analysis, 2008, 52(12): 5186-5201.

[11] HYNDMAN R J, SHANG H L. Rainbow plots, bagplots, and boxplots for functional data[J]. Journal of Computational and Graphical Statistics, 2010, 19(1): 29-45.

[12] 国家气象信息中心. 中国地面气候资料日值数据集(V3.0)[DB/OL]. [2018-03-01] http://data.cma.cn/data/cdcdetail/dataCode/SURF_CLI_CHN_MUL_DAY_V3.0.html.