Statistical and Application 统计学与应用, 2015, 4, 94-102
Published Online June 2015 in Hans. http://www.hanspub.org/journal/sa
http://dx.doi.org/10.12677/sa.2015.42011

Hans汉斯

Robust Scale Estimation Based on the Improved Median Absolute Deviations

Pingli Yang

China University of Mining and Technology (Beijing), Beijing
Email: xing122004@126.com

Received: Jun. 7th, 2015; accepted: Jun. 22nd, 2015; published: Jun. 29th, 2015

Abstract

Robust scale estimation with unknown location parameters which is called general median absolute deviations (*GMAD*) was proposed based on a robust scale estimation with location parameters of 0 (improved median absolute deviations FQ_n) given by Smirnor-Shevlyakov in 2014. The data analysis showed that FQ_n loses robustness when location parameters are unknown, but *GMAD* is robust when location parameters are zero or unknown.

Keywords

Scale Estimation, Robustness, Score Function

基于改进的中位数绝对偏差稳健尺度估计

杨苹莉

中国矿业大学(北京)，北京
Email: xing122004@126.com

收稿日期：2015年6月7日；录用日期：2015年6月22日；发布日期：2015年6月29日

摘　要

本文基于Smirnor-Shevlyakov在2014年针对位置参数已知为0的稳健尺度估计(即改进的中位数绝对偏

差FQ_n)，提出了位置参数未知时的稳健尺度估计(称之为广义中位数绝对偏差*GMAD*)。数据分析表明：FQ_n在位置参数未知时不稳健，但*GMAD*估计在位置参数为0以及未知时均稳健。

关键词

尺度估计，稳健性，得分函数

1. 引言

稳健性[1]考虑的是：当实际模型中的分布与假定模型中的分布有少许差异时，统计方法的性能会受到怎样的影响。因此，在粗差不可避免的情况下，选择适当的估计方法，使所估参数尽可能减免粗差的影响，得出正常模式下最佳或接近最佳的估值。

所谓的尺度参数是指满足分布族 $G(X)=F\left(\frac{x-b}{a}\right), a>0$，这里 a 便是尺度参数。

Huber 在 1981 年[2]就指出，在生产实践和科学实验所采集的数据中，粗差出现的概率为 1%~10%，并提到了一些高效稳健的尺度估计，如：四分位距(interquartile range) $IQR=F^{(-1)}(3/4)-F^{(-1)}(1/4)$，中位数绝对偏差(median absolute deviation) $MAD=med|x-\mu|$ (med 表示求中位数，下同)。*IQR* 估计非对称分布的尺度，它的崩溃点(breakdown point)最高可达 25%。*IQR* 针对的是对称分布，崩溃点最高可达 50%，高斯效为 37%，对于非对称分布则不适用。

Rousseuw-Croux 在 1933 年[3]依据四分位间距(0.25 quantile of the distances) $\{|x_i-x_j|\}$ 提出了两个更为高效的尺度估计量双中位数两两距离(double median of the pairwise distances) S_n，其表达式为：

$$S_n=cmed_i\{med_j|x_i-x_j|\}$$

其中 c 为纠偏因子。

下四分位两两距离(lower quartile of the pairwise distances) Q_n，表达式为：

$$Q_n=\{|x_i-x_j|,i<j\}_{(k)},\ k=\binom{h}{2},\ h=\left[\frac{n}{2}\right]+1,$$

它们的崩溃点均达到 50%，S_n 的高斯效可达到 58%，Q_n 可达到 82%。

Smirnor-Shevlyakov 在 2014 年[4]针对位置参数为 0 时，基于 Q_n 提出了改进的中位数绝对偏差(refinement of the median absolute deviation)，即为 FQ_n，其表达式为：

$$FQ_n=MAD^{(0)}*\left(1-\frac{U_0-\frac{n}{\sqrt{2}}}{U_2}\right),$$

其中

$$MAD^{(0)}=med|x|,$$

$$U_k=\sum_{i=1}^{n}u_i^k e^{-\frac{u_i^2}{2}},u_i=\frac{x_i}{MAD^{(0)}},k=0,2.$$

其计算速度为 Q_n 的 4~5 倍，崩溃点最高可达 50%，高斯效可达 80%，且更适用于蒙特卡罗模型。

2. 估计量 *GMAD*

基于 FQ_n 是针对刻度参数在均值为 0 时构造的估计，自然的想法就是将此估计推广至均值未知的场合。为此，我们构造如下的尺度估计量 *GMAD*：

$$GMAD^{(l)} = MAD * \left(1 - \frac{\left(6-\alpha^2\right)V_0 + \alpha^2 V_2 - \frac{12-\alpha^2}{2\sqrt{2}} n}{3\left(2-\alpha^2\right)V_2 + \alpha^2 V_4} \right),$$

其中 α 为自由参数，$l = 1,2$，l 取 1 时 $MAD = med\left|x - \bar{x}\right|$，$V_k = \sum_{i=1}^{n} v_i^k \mathrm{e}^{-\frac{v_i^2}{2}}$，$v_i = \frac{x_i - \bar{x}}{MAD}$，$k = 0,2,4$。

l 取 2 时，$MAD = med_i \left| x_i - med_j x_j \right|$，$V_k = \sum_{i=1}^{n} v_i^k \mathrm{e}^{-\frac{v_i^2}{2}}$，$v_i = \frac{x_i - med\left(x\right)}{MAD^{(2)}}$，$k = 0,2,4$。将 $GMAD^{(1)}$ 和 $GMAD^{(2)}$ 统称为 *GMAD*。

$MAD = med\left|x - \bar{x}\right|$，用样本均值 $\bar{x}$，估计 μ。而 $MAD = med_i \left| x_i - med_j x_j \right|$，用样本中位数，估计 μ。这两种估计量均是中位数绝对偏差的一种，在估计尺度参数时，对称分布中后者更稳健，在非对称分布中前者更稳健。

Huber 在 1981 年[2]提了 T 为估计量，F 为分布函数，T 在 F 处的影响函数为

$$IF\left(x;T,F\right) = \lim_{t\to 0} \frac{T\left(\left(1-t\right)F + t\Delta_x\right) - T\left(F\right)}{t},$$

其中 $x \in X$ 极限存在。

由于影响函数与得分函数程正相关，本文用得分函数来求。这里引用了参考文献[5]中的公式作为得分函数

$$\chi_a\left(x\right) = b - a^{-1}\left(\phi\left(x+a\right) - \phi\left(x-a\right)\right), \tag{1}$$

这里 b 是满足费希尔一致性条件 $\int \chi\left(x\right)\mathrm{d}\phi\left(x\right) = 0$ 的常数，$\alpha \geq 0$ 为自由参数，ϕ 是标准正态分布函数。

由泰勒展开式

$$\phi\left(x \pm \alpha\right) = \phi\left(x\right) \pm \alpha\varphi\left(x\right) + \frac{1}{2}\alpha^2\varphi'\left(x\right) \pm \frac{1}{6}\alpha^3\varphi''\left(x\right) + o\left(\alpha^3\right) \tag{2}$$

这里 φ 为标准正态分布密度函数。

由于 $\varphi = \phi'$，$\varphi' = -x\varphi$，$\varphi'' = \left(x^2 - 1\right)\varphi$，并将(2)代入(1)得

$$\chi_a\left(x\right) = b - \frac{1}{3}\left(6 + \alpha^2\left(x^2 - 1\right)\right)\varphi\left(x\right), \tag{3}$$

其中 $\alpha \geq 0$，由费希尔一致性条件 $\int \chi\left(x\right)\mathrm{d}\phi\left(x\right) = 0$，将(3)代入费希尔一致性条件，

$$0 = \int \chi\left(x\right)\mathrm{d}\phi\left(x\right) = b - \frac{1}{6\pi}\int\left(6 + \alpha^2\left(x^2 - 1\right)\right)\mathrm{e}^{-x^2}\mathrm{d}x,$$

由此得，

$$b = C\frac{12-\alpha^2}{12\sqrt{\pi}},$$

其中 C 为常数

由牛顿－辛普森一步迭代公式 $x_{k+1}=x_k-\frac{f(x_k)}{f'(x_k)}$ 知一步 M 估计

$$S_n^{(1)}=S_n^{(0)}-\frac{\sum\chi\left(\frac{x_i-\mu_i}{S}\right)}{\frac{\partial}{\partial S}\sum\chi\left(\frac{x_i-\mu_i}{S}\right)}$$

在 $S=S_n^{(0)}$ 处，其中 $S_n^{(0)}$ 为初始估计，所以

$$S_n^{(1)}=S_n^{(0)}\left(1+\frac{\sum\chi\left(\frac{x_i-\mu_i}{S_n^{(0)}}\right)}{\sum\chi\left(\frac{x_i-\mu_i}{S_n^{(0)}}\right)\chi'\left(\frac{x_i-\mu_i}{S_n^{(0)}}\right)}\right),$$

将(3)代入上式

$$S_n^{(1)}=S_n^{(0)}\left(1+\frac{\sum\chi\left(\frac{x_i-\mu_i}{S_n^{(0)}}\right)}{\sum\chi\left(\frac{x_i-\mu_i}{S_n^{(0)}}\right)\chi'\left(\frac{x_i-\mu_i}{S_n^{(0)}}\right)}\right)=S_n^{(0)}\left(1-\frac{\left(6-\alpha^2\right)U_0+\alpha^2U_2-\frac{12-\alpha^2}{2\sqrt{2}}n}{3\left(2-\alpha^2\right)U_2+\alpha^2U_4}\right),\tag{4}$$

其中

$$U_k=\sum_{i=1}^{n}u_i^k\mathrm{e}^{-\frac{u_i^2}{2}},\ u_i=\frac{x_i-\mu_i}{S_n^{(0)}},\ k=0,2,4.$$

用 $MAD=med\left|x-\bar{x}\right|$ 作为估计的初始值，μ 用 $\bar{x}$ 估计，代入(4)式，得

$$GMAD^{(1)}=MAD*\left(1-\frac{\left(6-\alpha^2\right)V_0+\alpha^2V_2-\frac{12-\alpha^2}{2\sqrt{2}}n}{3\left(2-\alpha^2\right)V_2+\alpha^2V_4}\right),$$

其中

$$V_k=\sum_{i=1}^{n}v_i^k\mathrm{e}^{-\frac{v_i^2}{2}},\ v_i=\frac{x_i-\bar{x}}{MAD},\ k=0,2,4.$$

用 $MAD^{(2)}=med_i\left|x_i-med_j x_j\right|$ 作为估计的初始值，μ 用 $med\left(x\right)$ 估计，代入(4)式，得

$$GMAD^{(2)}=MAD^{(2)}*\left(1-\frac{\left(6-\alpha^2\right)V_0+\alpha^2V_2-\frac{12-\alpha^2}{2\sqrt{2}}n}{3\left(2-\alpha^2\right)V_2+\alpha^2V_4}\right)$$

其中

$$V_k=\sum_{i=1}^{n}v_i^k\mathrm{e}^{-\frac{v_i^2}{2}},v_i=\frac{x_i-med\left(x\right)}{MAD^{(2)}},\ k=0,2,4.$$

$GMAD^{(1)}$ 和 $GMAD^{(2)}$ 统称为 $GMAD$。

由于 $\alpha \geq 0$ 为自由参数，不妨设 $\alpha = 0$，则 $GMAD$ 可简化为

$$GMAD = MAD * \left(1 - \frac{V_0 - \frac{n}{\sqrt{2}}}{V_2} \right).$$

3. 蒙特卡罗模拟

受污染分布描述为[1]：

$$P(X < x) = F(x-\theta) = (1-\varepsilon)\phi\left(\frac{x-\theta}{\sigma_0}\right) + \varepsilon\phi\left(\frac{x-\theta}{\sigma}\right),$$

其中 $\phi(\cdot)$ 表示标准正态分布函数，$\varepsilon > 0$ 是一个比较小的数，相应于异常观测值在全部观测值中所占的比例，而 σ^2 可能比 σ_0^2 大许多(或小许多)。由于粗差出现的概率为 1%~10%，故而实验中 ε 取 0.1。

3.1. 蒙特卡罗模拟结果图形

图 1~5 均为自由参数 $\alpha = 0$，$n = 20$，重复 10,000 次的结果。

图 1 为这几种尺度估计量在没有受到污染，且 $\sigma = 1$ 时的正态分布中变化情况图。由此图可知在没受到污染的正态分布中，SD 最接近 1，稳健性最好。其次是 $GMAD^{(1)}$ 和 $GMAD^{(2)}$，他们几乎是重合的，说明他们的估计尺度的稳健性几乎无差别。再者是 IQR 和 MAD，他们的估计稳健性也差不多。然后是 Q_n，虽有所偏离，但也在可承受范围内。我们能明显看出 FQ 的不同，在均值为 0 时，FQ 估计尺度的稳健性非常好，但在均值非 0 时，有明显的偏离，均值的绝对值越大，偏离程度越大，且偏离程度是对称的。

图 2 为各个尺度估计量在受污染的，$\sigma = 1$，$\sigma_* = 3\sigma$ 的正态分布中的变化情况图。由图 2 可看出，在受污染的正态分布中，$GMAD^{(1)}$ 与 $GMAD^{(2)}$ 估计尺度参数最接近 1，且它们几乎重合，说明它们估计尺度参数时的稳健性几乎差不多，且是这几个尺度估计量里估计尺度最好的。SD 估计尺度参数要比在没受污染的正态分布中要差些，其它的估计几乎与在没受污染的正态分布中估计几乎没什么差别。

由于伽马分布在形状参数 $0 < \alpha < 1$，$1 < \alpha < 2$，$\alpha > 2$ 时，对应的伽马密度曲线形状不同，故而在伽

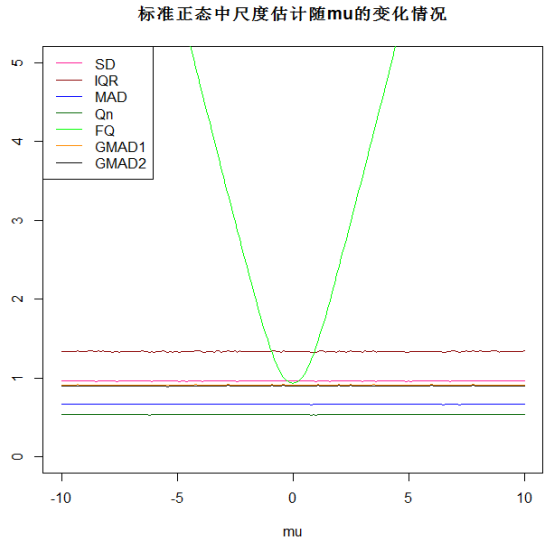

Figure 1. Scale estimators based on the mean of the standard normal distribution
图 1. 尺度估计量在标准正态分布中随均值变化图

马分布中的尺度估计量分这三种情况进行讨论，即图 3、图 4、图 5。

由图 3 可明显看出，这几种估计在形状参数 $0<\alpha<1$ 时，都不可行。

由图 4 可知，在形状参数 $1<\alpha<2$ 时，除了估计量 Q_n 的估计效果不佳外，其它的几个估计稳健性都挺好。*SD* 最接近 σ，稳健性最强，其次估计量 *IQR* 与新的估计量 $GMAD^{(1)}$ 稳健性也非常好，估计量 *FQ* 估计尺度时虽不如 $GMAD^{(1)}$，但也相对不错了，$GMAD^{(2)}$ 与 *MAD* 在此种情况下用来估计尺度也是可行的。

由图 5 可明显看出在形状参数 $\alpha>2$ 时，估计量 *FQ* 明显偏离准确尺度参数 σ 许多，说明此种估计根本不可行。估计量 *IQR* 也偏离准确尺度 2 倍，所以也不可行。其他的估计量都在允许范围内，且在这几种估计量估计尺度参数时，*MAD* 最为稳健，其次是 $GMAD^{(2)}$，然后是 $GMAD^{(1)}$ 与 Q_n，最后是 *SD*。

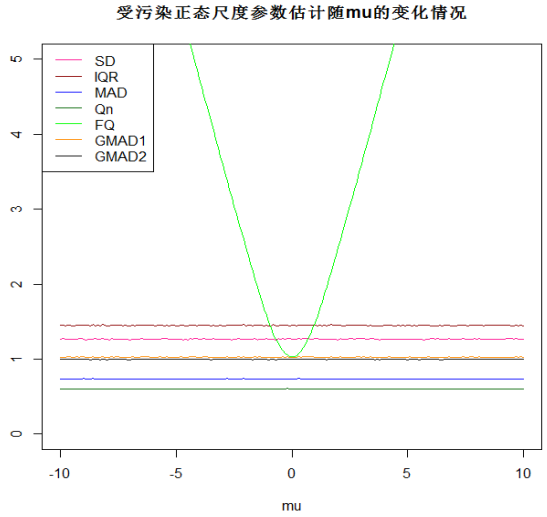

Figure 2. Scale estimators based on the mean of the contaminated normal distribution
图 2. 尺度估计量在受污染正态分布中随均值变化图

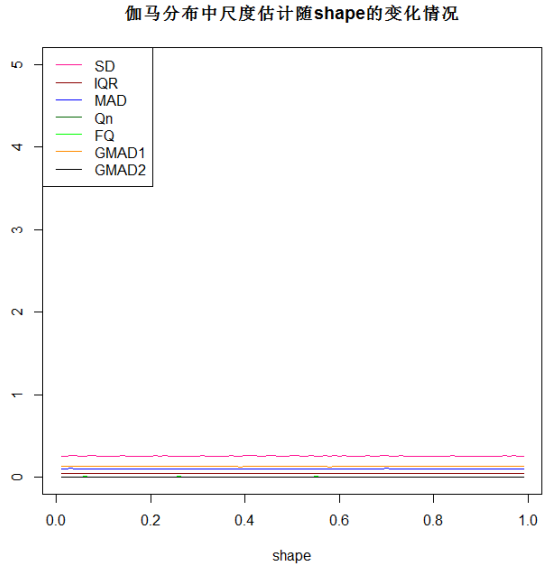

Figure 3. Scale estimation based on the gamma distribution of the shape parameter is from 0 to 1
图 3. 尺度估计量在形状参数为 0~1 的伽马分布中变化图

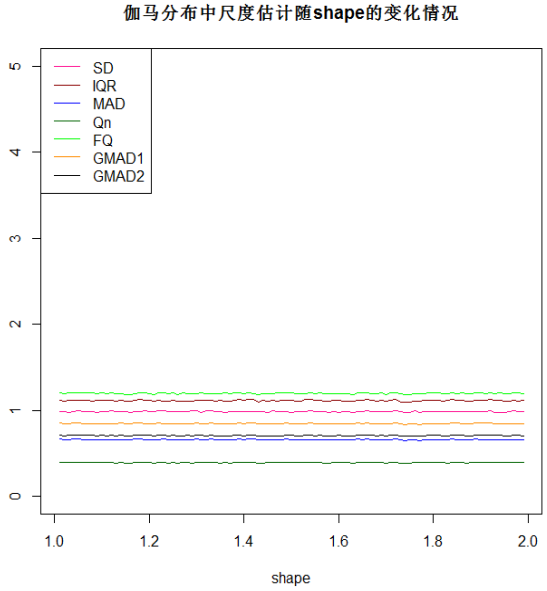

Figure 4. Scale estimators based on the Gamma distribution of the shape parameter is from 1 to 2
图 4. 尺度估计量在形状参数为 1~2 的伽马分布中变化图

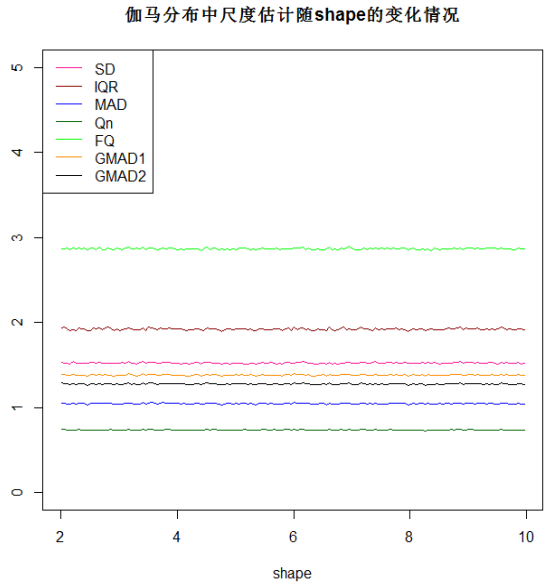

Figure 5. Scale estimators based on the gamma distribution of the shape parameter is greater than 2
图 5. 尺度估计量在形状参数大于 2 的伽马分布中变化图

由图我们可知，无论是在没有受到污染的正态分布中，还是在受到污染的正态分布中，新估计量 $GMAD^{(1)}$ 与 $GMAD^{(2)}$ 在估计尺度参数 σ 时，都非常稳健，而估计量 FQ 只有在均值为 0 时稳健。在伽马分布中，这几种估计量在形状参数 $0<\alpha<1$ 的情形下均不稳健；在形状参数 $1<\alpha<2$ 的情形下，除了估计量 Q_n 的稳健性不佳外，其它的几个估计量估计尺度参数时的稳健性都挺好；在形状参数 $\alpha>2$ 的情形下，估计量 FQ 与 IQR 非常不稳健，不能用来估计尺度参数，其它几个估计量均比较稳健。

3.2. 蒙特卡罗模拟结果表

表 1~3 均是自由参数 $\alpha = 0$，每个 n 都重复试验 10,000 次的结果。

由表 1 可知，在均值为 0 的受污染的正态分布中，估计量 FQ、$GMAD^{(1)}$、$GMAD^{(2)}$ 估计尺度参数时的稳健性几乎没什么差别，都很好。其它四中估计量估计尺度的偏差也都在可承受范围内。

表 2 为形状参数 $\alpha = 2$，尺度参数 $\lambda = 1$时的情况，此时估计量 FQ、$GMAD^{(1)}$、$GMAD^{(2)}$ 估计尺度参数时的稳健性都很好。其中 $GMAD^{(1)}$ 比 FQ 好些，FQ 比 $GMAD^{(2)}$ 好些。Q_n 超出了可承受范围。

表 3 说明在指数分布中，估计量 FQ 非常稳健，估计量 $GMAD^{(1)}$ 虽不如 FQ，但也很稳健。估计量 $GMAD^{(2)}$ 估计尺度参数的稳健性虽差些，但也在可承受范围内。

由表可知，对于均值为 0 时，在受污染的正态分布模型的情况下，FQ 与 $GMDA$ 的稳健性差不多。在非正态分布下的情形，如在形状参数为 2 时的伽马分布中，FQ 与 $GMDA$ 均可用，稳健性 FQ 虽不如 $GMAD^{(1)}$ 好，但比 $GMAD^{(2)}$ 要好。在指数分布中，估计量 FQ 非常稳健，估计量 $GMAD^{(1)}$ 虽不如 FQ，但也很稳健。估计量 $GMAD^{(2)}$ 估计尺度参数的稳健性虽差些，但也在可承受范围内。

Table 1. The mean of contaminated normal distribution is 0 ($\varepsilon = 10\%$, $\sigma = 1$, $\sigma_* = 3\sigma$)
表 1. 均值为 0 的受污染正态分布($\varepsilon = 10\%$, $\sigma = 1$, $\sigma_* = 3\sigma$)

n	Mean							Standardized variance						
	SD	IQR	MAD	Q_n	FQ	$GMAD^{(1)}$	$GMAD^{(2)}$	SD	IQR	MAD	Q_n	FQ	$GMAD^{(1)}$	$GMAD^{(2)}$
10	1.204	1.836	0.746	0.712	1.036	1.026	0.959	0.414	0.600	0.249	0.219	0.280	0.286	0.283
20	1.265	1.450	0.736	0.604	1.032	1.025	0.998	0.318	0.368	0.178	0.122	0.196	0.197	0.198
50	1.310	1.526	0.733	0.544	1.032	1.031	1.020	0.214	0.246	0.119	0.067	0.127	0.128	0.128
100	1.327	1.457	0.730	0.525	1.033	1.032	1.027	0.156	0.169	0.084	0.044	0.090	0.090	0.090

Table 2. Gamma distribution ($\alpha = 2, \lambda = 1$)
表 2. 伽马分布($\alpha = 2, \lambda = 1$)

n	Mean							Standardized variance						
	SD	IQR	MAD	Q_n	FQ	$GMAD^{(1)}$	$GMAD^{(2)}$	SD	IQR	MAD	Q_n	FQ	$GMAD^{(1)}$	$GMAD^{(2)}$
10	1.272	1.971	0.905	2.300	1.182	1.053	1.272	0.433	0.728	0.338	0.545	0.389	0.354	0.434
20	1.343	1.656	0.916	2.282	1.203	1.090	1.343	0.330	0.483	0.250	0.381	0.278	0.252	0.330
50	1.383	1.780	0.923	2.281	1.215	1.113	1.383	0.211	0.322	0.160	0.243	0.175	0.160	0.211
100	1.400	1.715	0.926	2.279	1.220	1.120	1.400	0.155	0.226	0.115	0.171	0.125	0.114	0.155

Table 3. Exponential distribution ($\lambda = 1$)
表 3. 指数分布($\lambda = 1$)

n	Mean							Standardized variance						
	SD	IQR	MAD	Q_n	FQ	$GMAD^{(1)}$	$GMAD^{(2)}$	SD	IQR	MAD	Q_n	FQ	$GMAD^{(1)}$	$GMAD^{(2)}$
10	0.873	1.216	0.608	0.439	1.081	0.779	0.627	0.359	0.529	0.251	0.181	0.385	0.298	0.253
20	0.936	1.038	0.624	0.358	1.069	0.797	0.650	0.280	0.341	0.183	0.098	0.271	0.213	0.182
50	0.970	1.124	0.629	0.314	1.060	0.801	0.662	0.188	0.233	0.117	0.053	0.172	0.136	0.117
100	0.984	1.085	0.632	0.301	1.061	0.804	0.667	0.134	0.162	0.084	0.035	0.122	0.096	0.083

4. 结论

由以上分析可知，正态分布模型无论有没有受到污染，新的估计量$GMAD^{(1)}$与$GMAD^{(2)}$在估计尺度参数σ时，都非常稳健，且它们的差异性不大，而估计量FQ只有在均值为0时才稳健。均值为正或负对估计的稳健性并没有影响，而均值的绝对值的大小对估计量FQ影响较大，对其它估计量也没什么影响。

在非正态分布下的情形，如伽马分布、指数分布等估计量FQ与新的估计量$GMAD$也是可用的。在线性模型中是否可用还有待研究。

致　谢

本文是在我的导师李再兴老师的指导完成的。此外，本文得到中央高校基本科研业务费以及北京市青年英才计划的资助。

参考文献 (References)

[1] 茆诗松, 等 (2006) 高等数理统计. 高等教育出版社, 北京, 147-156.

[2] Huber, P.J. (1981) Robust statistics. John Wiley & Sons, Inc., New York.

[3] Rousseeuw, P. and Croux, C. (1993) Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, **88**, 1273-1283.

[4] Smirnov, P.O. and Shevlyakov, G.L. (2014) Fast highly efficient and robust one-step M-estimators of scale based on Qn. *Computational Statistics & Data Analysis*, **78**, 153-158.

[5] Smirnov, P. and Shevlyakov, G. (2010) On approximation of the Qn-estimate of scale by fast M-estimates. In: *Book of Abstracts*: *International Conference on Robust Statistics*, *ICORS* 2010, Prague, Czech Republic, 94-95.