Robust correlations and outlier detection

Josué Cerpa Araña

6 de agosto de 2019

Shape indexes

Index MUOD shape

$$I_{S_3} = \left| \frac{1}{n} \sum_{j=1}^{n} \rho(x, x_j) - 1 \right|$$

Index tau kendall

$$I_{S_2} = \left| \frac{1}{n} \sum_{j=1}^{n} \tau(x, x_j) - 1 \right|$$

Index Spearman

$$I_{S_3} = \left| \frac{1}{n} \sum_{j=1}^{n} r(x, x_j) - 1 \right|$$

Bend correlation

https://www.itl.nist.gov/div898/software/dataplot/refman2/auxillar/pbendcr.htm

Winsorized correlation

The standard correlation estimate can be heavily influenced by extreme values. The Winsorized correlation compensates for this by setting the tail values equal to a certain percentile value. For example, for a 90% Winsorized correlation, the bottom 5% of the values are set equal to the value corresponding to the 5th percentile while the upper 5% of the values are set equal to the value corresponding to the 95th percentile. Then the standard correlation formula is applied.

Quadrant correlation

$$r_Q = \frac{1}{n} \sum_{i=1}^{n} sgn(x_i - med(x)) \, sgn(y_i - med(y))$$

Robust regression I

$$r_{REG} = \sqrt{\hat{\beta}_1 \hat{\beta}_2}$$

Median correlation coefficient

$$r_{MED} = \frac{med^2(|u|) - med^2(|v|)}{med^2(|u|) + med^2(|v|)}$$

where

$$u = \frac{x - med(x)}{\sqrt{2}\,MAD(x)} + \frac{y - med(y)}{\sqrt{2}\,MAD(y)}$$

$$v = \frac{x - med(x)}{\sqrt{2}\,MAD(x)} - \frac{y - med(y)}{\sqrt{2}\,MAD(y)}$$

MAD correlation coefficient

$$r_{MAD} = \frac{MAD^2(u) - MAD^2(v)}{MAD^2(u) + MAD^2(v)}$$

where

$$u = \frac{x - med(x)}{\sqrt{2}\,MAD(x)} + \frac{y - med(y)}{\sqrt{2}\,MAD(y)}$$

$$v = \frac{x - med(x)}{\sqrt{2}\,MAD(x)} - \frac{y - med(y)}{\sqrt{2}MAD(y)}$$

Biweight mid correlation

https://www.itl.nist.gov/div898/software/dataplot/refman2/auxillar/biwmidc.htm

```
## Warning in min(x): ningún argumento finito para min; retornando Inf

## Warning in max(x): ningun argumento finito para max; retornando -Inf

## Warning in min(x): ningún argumento finito para min; retornando Inf

## Warning in max(x): ningun argumento finito para max; retornando -Inf

## $MEASURES.PEARSON
##                   c         f         F
## boxplot    98.46699 2.5032232 0.7839797
## adjboxplot 32.00213 0.0000000 0.4067427
## adjout      0.00000 0.0000000 0.0000000
## meanvar    83.47244 0.1536069 0.8753619
## carlings   98.13763 2.2204622 0.7990245
## madmedian  99.52652 6.4660508 0.6216892
## ifourths   98.21699 2.4612276 0.7847660
## tangent    94.67363 1.8664015 0.8126934
##
## $MEASURES.KENDALL
##                   c          f         F
## boxplot    69.16566   4.987359 0.5878899
## adjboxplot 23.99048   1.524632 0.1860611
## adjout      0.00000   0.000000 0.0000000
## meanvar    80.40898   0.000000 0.8789873
## carlings   71.74982   4.987359 0.6152862
## madmedian  97.92063  18.793295 0.3979751
```

```
## ifourths    68.16566  4.987359 0.5816475
## tangent     99.33333 19.359109 0.4276143
##
## $MEASURES.SPEARMAN
##                   c          f         F
## boxplot    81.92688  4.987359 0.6967844
## adjboxplot 25.35595  2.421586 0.1622914
## adjout      0.00000  0.000000 0.0000000
## meanvar    78.94037  0.000000 0.8690017
## carlings   86.52330  4.987359 0.7279081
## madmedian  99.55556 21.821817 0.3339590
## ifourths   81.76021  4.987359 0.6958753
## tangent    96.23413 13.255820 0.5779607
##
## $MEASURES.BEND
##                   c          f         F
## boxplot    97.12255 2.5242759 0.7753200
## adjboxplot 26.83384 0.0000000 0.3482570
## adjout      0.00000 0.0000000 0.0000000
## meanvar    81.84271 0.2053725 0.8590283
## carlings   96.76144 2.2520573 0.7895906
## madmedian  99.35985 6.6550702 0.6144272
## ifourths   97.12255 2.4822803 0.7777576
## tangent    91.71053 1.9049821 0.7903342
##
## $MEASURES.WINSOR
##                   c          f         F
## boxplot    97.44755 2.5140856 0.7776700
## adjboxplot 26.97071 0.0000000 0.3504356
## adjout      0.00000 0.0000000 0.0000000
## meanvar    81.15065 0.2053725 0.8529561
## carlings   96.99358 2.2314365 0.7920782
## madmedian  99.24874 6.5083964 0.6184208
## ifourths   97.19358 2.4717539 0.7785047
## tangent    92.05339 2.0413397 0.7888944
##
## $MEASURES.QUADRANT
##                   c          f         F
## boxplot    51.87478  7.9735086 0.4019924
## adjboxplot 18.53135  2.4381256 0.1372527
## adjout      0.00000  0.0000000 0.0000000
## meanvar    38.44026  0.4377296 0.5015497
## carlings   51.13193  7.8327557 0.4111840
## madmedian  68.80184 21.1024133 0.2446772
## ifourths   51.54145  7.8327557 0.4029924
## tangent    61.45851 14.3212038 0.3608166
##
## $MEASURES.ROBUSTREG
##                   c          f         F
## boxplot    98.46699 2.5032232 0.7839797
## adjboxplot 32.00213 0.0000000 0.4067427
## adjout      0.00000 0.0000000 0.0000000
## meanvar    83.47244 0.1536069 0.8753619
## carlings   98.13763 2.2204622 0.7990245
```

```
## madmedian  99.52652 6.4660508 0.6216892
## ifourths    98.21699 2.4612276 0.7847660
## tangent     94.67363 1.8664015 0.8126934
##
## $MEASURES.MEDIANCOR
##                    c         f         F
## boxplot    66.76252 0.8914663 0.6882768
## adjboxplot 15.98662 0.0000000 0.2240771
## adjout      0.00000 0.0000000 0.0000000
## meanvar    59.47594 0.4233583 0.6745649
## carlings   63.84304 0.7641606 0.6798565
## madmedian  76.04423 3.6625199 0.6167121
## ifourths   66.76252 0.8706307 0.6898244
## tangent    79.75137 4.5635346 0.6236258
##
## $MEASURES.MADCOR
##                     c        f         F
## boxplot    63.678427 3.134125 0.5362075
## adjboxplot  8.629004 0.000000 0.1306674
## adjout      0.000000 0.000000 0.0000000
## meanvar    51.967208 1.016876 0.5712687
## carlings   62.748629 2.881844 0.5401826
## madmedian  80.955014 8.544887 0.4633313
## ifourths   63.678427 3.061306 0.5399086
## tangent    69.641017 5.759713 0.5174606
##
## $MEASURES.BIWEIGHT
##                     c          f          F
## boxplot    96.711833  1.00313617 0.88308207
## adjboxplot  7.311977  0.00000000 0.09928327
## adjout      0.000000  0.00000000 0.00000000
## meanvar    79.868074  0.08153171 0.86018117
## carlings   97.354690  1.10676456 0.87418370
## madmedian  99.582071 11.20044956 0.51178407
## ifourths   96.711833  0.99293209 0.88344571
## tangent    92.788817  1.52582029 0.82425205
```