Introduction to Mazeinda

Alice Dolma Albasi, Martin Wells

2018-01-15

Introduction

Mazeinda provides functions for detecting and testing the significance of pairwise Monotonic Association for Zero-Inflated non-negative Data with the measure proposed by Pimentel(2009)[1]. This package can handle any degree of zero inflation as well as non-zero ties. The unique requirement is the independence of the entries of the vectors to be correlated. Pairwise association and testing of a large number of vectors can be expedited with parallel computations thanks to the packages "foreach" and "doMC".

A development of τ^*

In the setting of zero inflated data Pimentel(2009) suggested that τ^* can be re-expressed as

$$\tau^* = p_{11}^2 \tau_{11} + 2(p_{00}p_{11} - p_{01}p_{10}) \tag{1}$$

where

- p_{00} is the probability that for any given pair of corresponding entries of X and Y both entries are 0;
- p_{11} is the probability that for any given pair of corresponding entries of X and Y both entries are positive;
- p_{10} is the probability that for any given pair of corresponding entries of X and Y the entry of the vector X is positive while the entry of the vector Y is zero;
- p_{01} is the probability that for any given pair of corresponding entries of X and Y the entry of the vector X is zero while the entry of the vector Y is positive; and
- τ_{11} is the value of Kendall τ_b computed on the strictly positive entries of the vectors X and Y.

Recall that two pairs of jointly distributed non-negative continuous random variables (X_i, Y_i) and (X_j, Y_j) are said to be concordant if $(X_i - X_j)(Y_i - Y_j) > 0$ and discordant if $(X_i - X_j)(Y_i - Y_j) < 0$. There are 16 possible different realizations of two pairs of the realization of the random variables (x_i, y_i) and (x_j, y_j) displayed in Table 1 below.

To calculate the probability of concordance we apply the law of total probability using the two left most columns of Table 1, it follows that

$$\begin{aligned} P(C) &= \sum_{\text{all pairs}} P[C|\, X_i, Y_i, X_j, Y_j]\; P[X_i, Y_i, X_j, Y_j] \\ &= 2p_{00}p_{11} + p_{01}p_{11}P(Y_i < Y_j) + p_{10}p_{11}P(X_i < X_j) \\ &\qquad + p_{01}p_{11}P(Y_i > Y_j) + p_{10}p_{11}P(X_i > X_j) \\ &\qquad + p_{11}^2 P[(X_i - X_j)(Y_i - Y_j) > 0]. \end{aligned} \tag{2}$$

Since for positive values of the continuous random variables X and Y we have $P(X_i < X_j) = 1 - P(X_i > X_j)$ and $P(Y_i < Y_j) = 1 - P(Y_i > Y_j)$, (2) reduces to

$$P(C) = 2p_{00}p_{11} + p_{01}p_{11} + p_{10}p_{11} + p_{11}^2 P[(X_i - X_j)(Y_i - Y_j) > 0]. \tag{3}$$

Table 1: All Possible Realization and Conditional Probabilities of Concordance and Discordance

$P[X_i, Y_i, X_j, Y_j]$	$P[C \vert\ X_i, Y_i, X_j, Y_j]$	$P[D \vert\ X_i, Y_i, X_j, Y_j]$
$P[0,0,0,0] = p_{00}^2$	0	0
$P[1,0,0,0] = p_{00}p_{10}$	0	0
$P[0,1,0,0] = p_{00}p_{01}$	0	0
$P[0,0,1,0] = p_{00}p_{10}$	0	0
$P[0,0,0,1] = p_{00}p_{01}$	0	0
$P[1,1,0,0] = p_{00}p_{11}$	1	0
$P[0,1,1,0] = p_{01}p_{10}$	0	1
$P[0,0,1,1] = p_{00}p_{11}$	1	0
$P[1,0,0,1] = p_{01}p_{10}$	0	1
$P[1,0,1,0] = p_{10}^2$	0	0
$P[0,1,0,1] = p_{01}^2$	0	0
$P[0,1,1,1] = p_{01}p_{11}$	$P(Y_i < Y_j)$	$P(Y_i > Y_j)$
$P[1,0,1,1] = p_{10}p_{11}$	$P(X_i < X_j)$	$P(X_i > X_j)$
$P[1,1,0,1] = p_{01}p_{11}$	$P(Y_i > Y_j)$	$P(Y_i < Y_j)$
$P[1,1,1,0] = p_{10}p_{11}$	$P(X_i > X_j)$	$P(X_i < X_j)$
$P[1,1,1,1] = p_{11}^2$	$P[(X_i - X_j)(Y_i - Y_j) > 0]$	$P[(X_i - X_j)(Y_i - Y_j) < 0]$

The same logic can be applied for the calculation the probability of discordance using the first and third columns of Table 1 to show that

$$P(D) = 2p_{01}p_{10} + p_{01}p_{11} + p_{10}p_{11} + p_{11}^2 P[(X_i - X_j)(Y_i - Y_j) < 0]. \tag{4}$$

Finally, Kendall's tau formula is the difference of (3) and (4) so that

$$\tau^* = 2p_{00}p_{11} - 2p_{01}p_{10} + p_{11}^2 \tau_{11}.$$

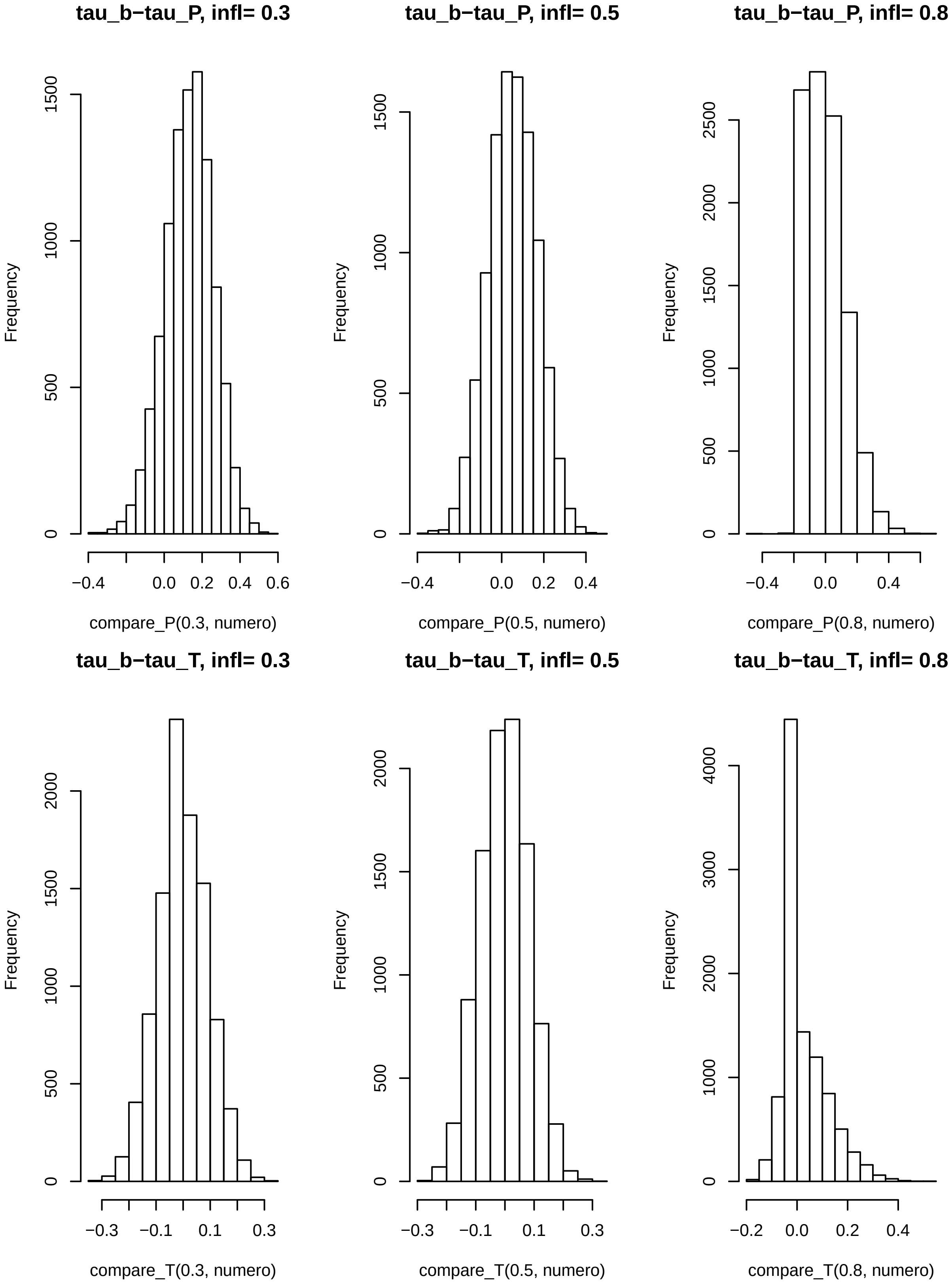
Simulations

The choice of the measure proposed in Pimentel(2009)[1] over the one by Pimentel, Niewiadomska-Bugajb and Wang(2015) [3] is motivated by the proof above and by the following simulations which suggest that the method considered in this package has superior performance in terms of bias and power.

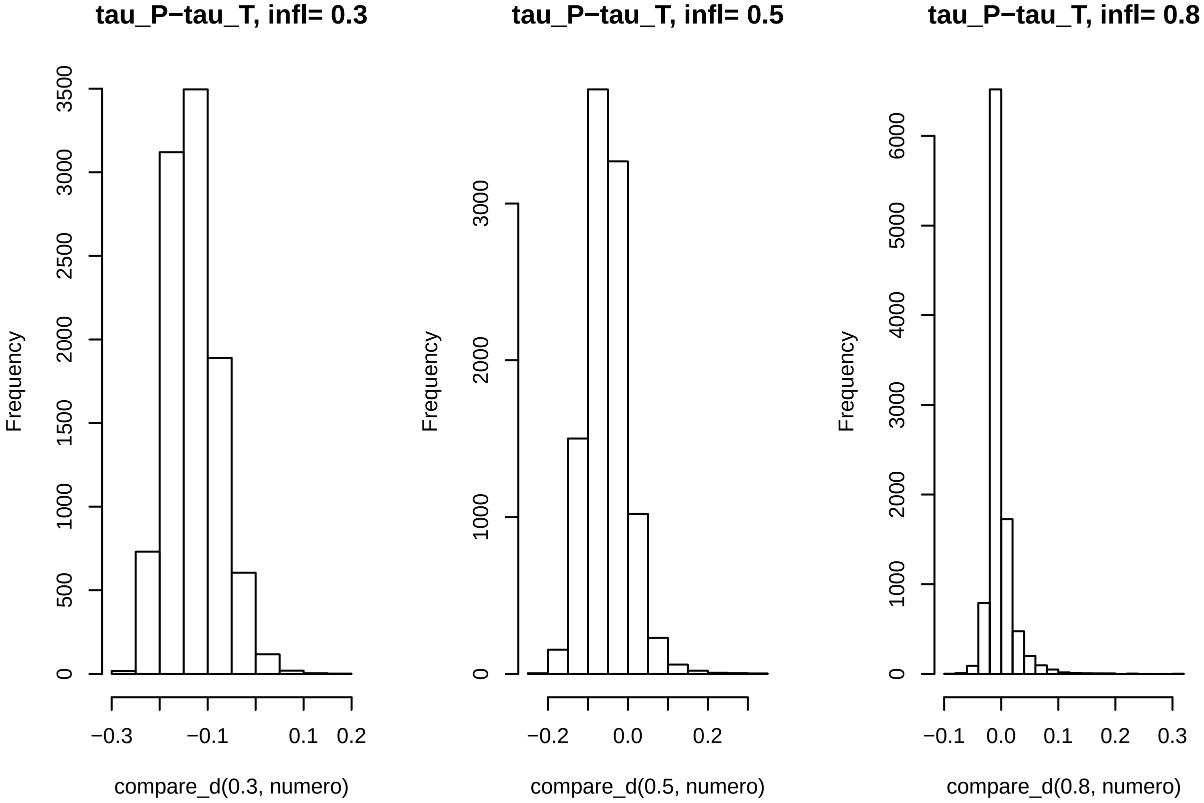
Let:

- τ_b be Kendall's τ_b
- τ_t= $p_{11}^2 t_{11} + 2(p_{00}p_{11} - p_{01}p_{10})$
- τ_p= $p_{11}^2 t_{11} + 2(p_{00}p_{11} - p_{01}p_{10}) + 2p_{11}[p_{10}(1 - 2p_1) + p_{01}(1 - 2p_2)]$
- d= $\tau_t - \tau_p$= $2p_{11}[p_{10}(1 - 2p_1) + p_{01}(1 - 2p_2)]$

The zero-inflated data is derived from a beta zero-inflated distribution with parameters $(0.5, 1)$ and probability mass at 0 of 0.3, 0.5 and 0.8. The first figure below shows the histograms of $\tau_b - \tau_p$ and the following the histogram of $\tau_b - \tau_t$ obtained with 10.000 vectors with 30 entries.



While in the histograms for 0.8 inflation τ_t and τ_p seem to behave similarly and the mode in the histogram for 0.8 of the difference d is close to zero, the histograms for 0.3 and 0.5 inflation suggest that τ_t is a better estimator for Kendall's τ_b as the mode is centered at 0.



The means of the values plotted in the histograms above confirm these results:

Table 2: means of the two statistics

	tau_p	tau_t
infl=0.3	0.1318665	-0.0007732
infl=0.5	0.0531340	-0.0000680
infl=0.8	0.0005677	0.0334389

Comments about $\tau*$ and code

The formula for the variance of $\tau*$ proposed in Pimentel's thesis [1] cannot be applied when the parameters p_{ij} attain the values 0 or 1: the proof relies on the delta method which can be applied only when functions are differentiable. At the extrema of the interval [0,1] no function is differentiable, therefore, whenever any of the parameters p_{ij} attains the values 1 or 0, the standard R *cor.test* function for Kendall's correlation is used to obtain p-values, when applicable, as the function does not work in case of vectors with zero variance.

In case of non-zero ties, τ_{11} is calculated as Kendall's τ_b. If all the non-zero entries of at least one vector are tied, τ_{11} os set to 0. This choice is in accordance with the treatment of ties in Kendall (1970)[2]. τ_{11} is set to 0 also in the case of only one pair of non-zero corresponding entries.

Examples

Mazeinda provides for the user three functions: *associate* for obtaining a matrix of association values, *test_associations* for obtaining the p-values of the output of associate and *combine* which gives non-zero values only if the association is significantly different from 0. It should be noted that independence of two variates implies a zero value of $\tau*$, but it is not a sufficient and necessary condition and this affects the power of the test, as dependent cases might get rejected. All these functions can be called with two vectors as arguments or with two matrices with vectors to be correlated as columns. If the vectors to be correlated belong to one matrix m, m should be specified as the argument m1 and as the argument m2.

Note that if the parameters p_{ij} attain the values 0 or 1, since the R function *cor.test* must be used, the function *combine* reports the value of Kendall's τ_b given by the stardart R *cor* function, if significantly different from 0. The R *cor* cannot handle vectors with standard deviation 0, therefore such vectors will yield NA. Below are some examples; warnings are suppressed because the *cor.test* function prints "cannot compute exact p-value with ties".

```r
library(mazeinda)

set.seed(234)

x=abs(rBEZI(50, mu = 0.9, sigma = 1, nu = 0.7))
y=abs(rBEZI(50, mu = 0.9, sigma = 1, nu = 0.2))
associate(x,y)
```

```
## [1] -0.0256
```

```r
test_associations(x,y)
```

```
## [1] 0.6678331
```

```r
combine(x,y)
```

```
## [1] 0
```

```r
x=matrix(abs(rBEZI(50, mu = 0.9, sigma = 1, nu = 0.6)),5)
y=matrix(abs(rBEZI(30, mu = 0.9, sigma = 1, nu = 0.3)),5)
knitr::kable(associate(x,y), digits=3, caption="associate(x,y)")
```

Table 3: associate(x,y)

0.120	-0.756	0.080	0.080	-0.120	-0.120
0.316	-0.500	0.378	-0.378	0.000	-0.632
0.527	0.240	-0.080	-0.080	0.316	-0.527
0.000	0.667	-0.378	-0.378	0.316	0.000
-0.120	-0.080	0.714	-0.571	-0.359	-0.837
0.359	-0.080	-0.571	1.000	0.359	0.598
0.359	0.630	-0.571	0.080	0.598	0.120
0.105	-0.444	-0.080	-0.080	-0.105	-0.105
-0.600	-0.316	0.359	-0.120	-0.800	0.000
0.316	-0.500	0.378	-0.378	0.000	-0.632

```r
knitr::kable(suppressWarnings(test_associations(x,y)),
             digits=3, caption="test_associations(x,y)")
```

Table 4: test_associations(x,y)

0.782	0.087	0.355	0.355	0.782	0.782
0.480	0.277	0.429	0.429	1.000	0.157
0.207	0.232	0.645	0.645	0.448	0.207
1.000	0.147	0.429	0.429	0.480	1.000
0.782	0.645	0.119	0.213	0.405	0.052
0.405	0.645	0.213	0.029	0.405	0.166
0.405	0.154	0.213	0.355	0.166	0.782
0.801	0.298	0.645	0.645	0.801	0.801
0.233	0.448	0.405	0.782	0.083	1.000
0.480	0.277	0.429	0.429	1.000	0.157

```r
knitr::kable(suppressWarnings(combine(x,y)),
             digits=3,caption="combine(x,y)")
```

Table 5: combine(x,y)

0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	1	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0

```r
knitr::kable(suppressWarnings(associate(x,x)),
             digits=3, caption="associate all vectors in the matrix x pairwise")
```

Table 6: associate all vectors in the matrix x pairwise

1.000	0.756	-0.080	-0.378	0.080	0.080	-0.571	0.882	0.359	0.756
0.756	1.000	0.667	-0.250	0.756	-0.378	-0.378	0.667	0.000	1.000
-0.080	0.667	1.000	0.333	-0.080	-0.080	0.378	0.240	-0.316	0.667
-0.378	-0.250	0.333	1.000	-0.378	-0.378	0.756	0.000	-0.316	-0.250
0.080	0.756	-0.080	-0.378	1.000	-0.571	-0.571	-0.080	0.120	0.756
0.080	-0.378	-0.080	-0.378	-0.571	1.000	0.080	-0.080	-0.120	-0.378
-0.571	-0.378	0.378	0.756	-0.571	0.080	1.000	-0.080	-0.598	-0.378
0.882	0.667	0.240	0.000	-0.080	-0.080	-0.080	1.000	0.316	0.667
0.359	0.000	-0.316	-0.316	0.120	-0.120	-0.598	0.316	1.000	0.000
0.756	1.000	0.667	-0.250	0.756	-0.378	-0.378	0.667	0.000	1.000

Computations in parallel

If the number of vectors to be associated is extensive, computations can be done in parallel for any of the functions in the package by specifying “parallel=TRUE” and “n_cor=n”, where n is the number for cores desired.

The parameters p_{ij}’s necessary for calculating $\tau*$ are by default calculated as proportions based on the entries of each pair of vectors to be associated. If the parameters are known, they can be provided to any of the functions in the package by specifying the values as arguments. Only p_{11}, p_{01} and p_{10} can be specified. Note that if one of these parameters is specified as an argument, the other two must be specified as well. The parameters p_{ij}'s can be estimated as by the population mean, calculated using all the pairs of column vectors from the matrices specified as the arguments d1 and d2.

```
knitr::kable(associate(x,y, estimator="own", p11=0.1,p10=0.1, p01=0.1), digits=3,
             caption="parameters specified")
```

Table 7: parameters specified

0.130	0.12	0.12	0.12	0.130	0.120
0.120	0.12	0.12	0.12	0.120	0.120
0.117	0.13	0.12	0.12	0.110	0.110
0.120	0.12	0.12	0.12	0.120	0.120
0.120	0.12	0.11	0.12	0.120	0.120
0.130	0.12	0.12	0.13	0.130	0.110
0.110	0.13	0.12	0.12	0.110	0.110
0.123	0.12	0.12	0.12	0.117	0.130
0.117	0.11	0.13	0.12	0.110	0.123
0.120	0.12	0.12	0.12	0.120	0.120

```
m1=matrix(abs(rBEZI(50, mu = 0.9, sigma = 1, nu = 0.7)),5)
m2=matrix(abs(rBEZI(30, mu = 0.9, sigma = 1, nu = 0.3)),5)
knitr::kable(associate(x,y, estimator="mean", d1=m1,d2=m2), digits=3,
             caption="parameters estimated as population mean")
```

Table 8: parameters estimated as population mean

0.044	0.000	0.000	0.000	0.044	0.000
0.000	0.000	0.000	0.000	0.000	0.000
-0.015	0.044	0.000	0.000	-0.044	-0.044
0.000	0.000	0.000	0.000	0.000	0.000
0.000	0.000	-0.044	0.000	0.000	0.000
0.044	0.000	0.000	0.044	0.044	-0.044
-0.044	0.044	0.000	0.000	-0.044	-0.044
0.015	0.000	0.000	0.000	-0.015	0.044
-0.015	-0.044	0.044	0.000	-0.044	0.015
0.000	0.000	0.000	0.000	0.000	0.000

References

[1] Pimentel, R. S. (2009). Kendall's Tau and Spearman's Rho for Zero-Inflated Data. Western Michigan University Dissertation 721 `http://scholarworks.wmich.edu/dissertations/721`.

[2] Kendall, G.M. (1970). *Rank Correlation Methods.* 4th ed. Griffin, London.

[3] Pimentel, R.S., Niewiadomska-Bugajb, M., and Wang, J. (2015) Association of zero-inflated continuous variable. *Statistics & Probability Letters* 96, 61-67.