Evaluating the Quality of Correlation Estimates Under Data Point Censoring

Fitsum A. Ayele[1] and Kimberly A. Barchard[1]

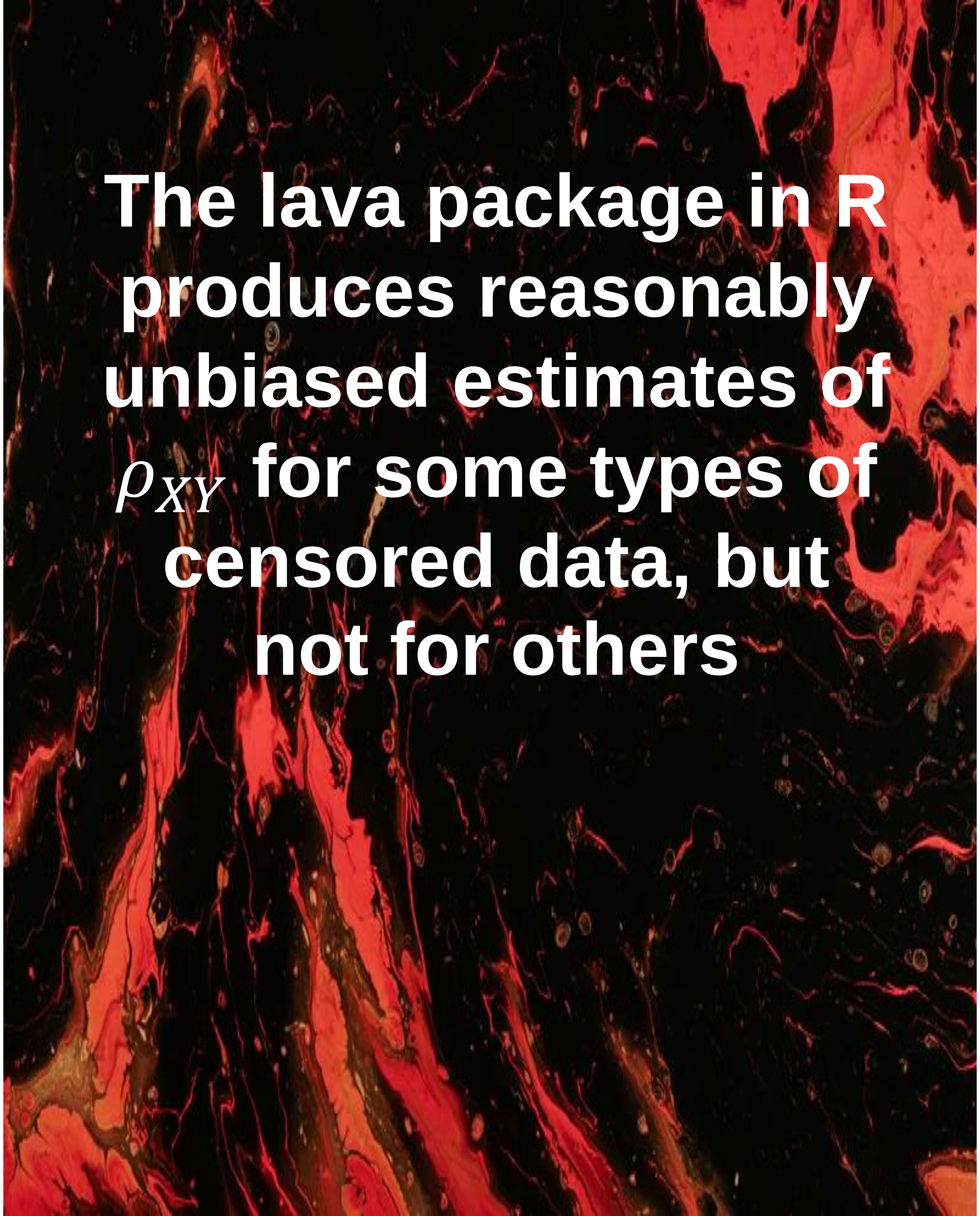
1. Department of Psychology, University of Nevada, Las Vegas

Introduction

- Oftentimes data in many settings are censored. When a data point is censored, it means that we only have partial information on its true value, knowing the value is no larger than, or at least as large as, a limit of detection (Fox, 2016).
- Censoring can skew results and discredit conclusions (Barchard & Russell, 2020).
- Fortunately, there a vast array of tools to correct the effect of censoring when estimating single variable statistics like medians and means.
- However, few methods exist that deal with correlations when both variables have been censored. Correlations are used widely throughout psychology to examine a plethora of relationships.
- Therefore, the purpose of this paper is to evaluate a method for estimating the correlation between two variables when they are both censored.
- We examined the *lava* package in R, and assessed its ability to estimate ρ_{XY} (Holst et al., 2015).
- In order to test out lava's ability to estimate ρ_{XY}, we calculated bias, which was the mean difference between the estimates ($\hat{\rho}_{XY}$) and the true values(ρ_{XY}).
- In order to evaluate if *lava's* estimates are accurate, we used a Monte Carlo study. We aimed to investigate if estimates of ρ_{XY} are biased when x and y are censored, and to assess how that bias varies for different values of ρ_{XY}, and different patterns of censoring.

Method

- This study used a total of 40 cells. Each cell was a unique combination of sample size, ρ_{XY}, and censoring pattern.
- We used a sample size of 500. We selected a large sample size in order to provide *lava* with ample information to create the best estimates of ρ_{XY} that it can.
- We used 10 values of ρ_{XY}: -.10, -.30, .30, -.50, .50, -.65, -.80, .80, -.95, and .95. A majority (six) of the values we chose were negative. We did this because when both variables are censored in the same direction, there is a greater effect on negative relationships.
- We used four different censoring patterns on x and y. Patterns that involved 20%, 40%, 75%, and 90% censoring on x and y, indicating low, moderate, heavy, and extremely heavy censorship, respectively.
- We ran 1000 trials for each cell. Within each trial, we produced a random set of data for which X and Y had a multivariate normal distribution with the selected correlation, and we censored *x* and *y* to the required percentages. We provided *lava* the *x* and *y* data. We then asked *lava* to estimate the correlation between X and Y (ρ_{XY}). The estimate $\hat{\rho}_{XY}$ used maximum likelihood estimation based upon the correlation model (Holst, 2020). In order to assess the quality of the estimates of ρ_{XY}, we calculated bias (the mean difference between $\hat{\rho}_{XY}$ and ρ_{XY}). Ideally, bias should be low (close to 0).



The lava package in R produces reasonably unbiased estimates of ρ_{XY} for some types of censored data, but not for others

Results

- We assembled the results of the Monte Carlo simulation into a table. See Table 1.
- Overall, we found that there was very little bias for positive values of ρ_{XY}. In addition, there was very little bias for 20% and 40% censoring across all values of ρ_{XY}.
- However, when it came to 75% and 90% censoring, and negative values of ρ_{XY}, there was substantial bias.
- An interesting finding is that for 90% censoring, if ρ_{XY} was not close to -1, the bias was negative and the estimates were closer to -1. However, if ρ_{XY} was really close to -1 (as in the value of -.95 used in this study), the bias was positive and the estimates were further from -1.
- This trend was not observed with 75% censoring, where the bias was always negative for negative values of ρ_{XY}, producing estimates closer to -1 than the true value of ρ_{XY}.

Table 1

Bias for Lava Estimates of ρ_{XY} for Different Values of ρ_{XY} Under Different Patterns of Censoring

	Estimates of Bias for Different Patterns of Censoring			
ρ_{XY}	20 %	40%	75%	90%
-0.95	.00	.00	-.02	.03
-0.80	.00	.00	-.06	-.12
-0.65	.00	.00	-.01	-.24
-0.50	.00	.00	-.01	-.25
-0.30	.00	.00	.00	-.11
-0.10	.00	.00	.00	-.02
0.95	.00	.00	.00	.00
0.80	.00	.00	.00	.00
0.50	.00	.00	.00	-.01
0.30	.00	.00	.00	-.01

Note. Percentages of censoring are the same on x and y.

Discussion

- *Lava* produces accurate estimates of ρ_{XY} for positive to moderately negative values of ρ_{XY} and low degree of censoring; but, for strong negative values of ρ_{XY} and high degree of censoring, it is not as accurate.
- If researchers use this method with this type of data, their estimates will be biased, leading to inaccurate conclusions.
- In particular, when researchers are interested in large negative correlations, they should attempt to have minimal censoring on their data points in order to avoid substantial bias.
- For example, if right censoring is caused by participants not experiencing the event before the study ends, researchers may wish to extend their study to avoid an abundance of right censored points.
- In addition, in order to reduce left censored values, researchers can use more sensitive measuring instruments with smaller lower levels of detection, leading to a minimization of left censored points in the dataset.
- Future research can assess lava for non-normal distributions.

UNLV