Compare permutation and scalar transformation approaches on simulated and real data

Sayani Gupta

08/02/2021

Simulated data generation: Observations are generated from a N(0,1) distribution for each combination of nx and $nfacet$ from the following sets: $nx = nfacet = \{2, 3, 5, 7, 14, 20, 31, 50\}$ to cover a wide range of levels from very low to moderately high. Each combination is being referred to as a *panel*. That is, data is being generated for each of the panels $\{nx = 2, nfacet = 2\}, \{nx = 2, nfacet = 3\}, \{nx = 2, nfacet = 5\}, \ldots, \{nx = 50, nfacet = 31\}, \{nx = 50, nfacet = 50\}$. For each of the 64 panels, $ntimes = 500$ observations are drawn for each combination of the categories. That is, if we consider the panel $\{nx = 2, nfacet = 2\}$, 500 observations are generated for each of the combination of categories from the panel, namely, $\{(1, 1), (1, 2), (2, 1), (2, 2)\}$. The values of λ is set to 0.67 and values of raw wpd wpd_{raw} is obtained.

Scalar transformation approach to normalisation

A log-linear model is fitted to see how the values of wpd_{raw} changes with the values of nx and $nfacet$. The model is of the form

$$y = a + b * log(x) + e$$

, where $y = median(wpd_{raw})$ and $x = nx * nfacet$. wpd_norm is a transformation on wpd_{raw} which should be designed to remove the effect of $nx * nfacet$ on wpd_{raw} and thus is defined as follows: $wpd_{norm} = wpd_{raw} - b * log(nx * nfacet)$

Permutation approach to normalisation

The simulated data for each of the panels is permuted/shuffled $nperm = 100$ times and for each of those permutations wpd_{norm} is computed as follows: $wpd_{norm} = (wpd_{raw} - mean(wpd_{raw}))/sd(wpd_{raw})$. This is done so that the distribution of the normalised measure wpd_{norm} has the same mean and standard deviation across different nx and nfacet.