Publication Supplement

Institute of Bioinformatics, Johannes Kepler University Linz

Testing Noisy Numerical Data for Monotonic Association

— supplementary material —

Ulrich Bodenhofer[1], Martin Krone[2], and Frank Klawonn[2,3]

[1]Institute of Bioinformatics, Johannes Kepler University, Linz, Austria
[2]Department of Computer Science, Ostfalia University of Applied Sciences, 38302 Wolfenbüttel, Germany
[3]Bioinformatics and Statistics, Helmholtz Centre for Infection Research, 38124 Braunschweig, Germany

S1 An explanation for the limits in Table 1

In case of the truncated linear scoring or the ε-tolerant strict ordering, one can derive explicit expressions for the overall scores of concordant pairs $\tilde{C}_{n,k}$ and discordant pairs $\tilde{D}_{n,k}$ if $k < 0.25n$ outliers that result in the largest possible decrease of $\tilde{\gamma}$ are present. In these cases, exact values of the one-sided limit

$$\lim_{a \to 0.25^-} \left(\lim_{\substack{n\to\infty \\ k=\lfloor a \cdot n \rfloor}} \frac{\tilde{C}_{n,k} - \tilde{D}_{n,k}}{\tilde{C}_{n,k} + \tilde{D}_{n,k}} \right)$$

were obtained. Using the example of the truncated linear scoring and the minimum t-norm $T_{\mathbf{M}}$, we show how these limits can be derived if the parameter of the ordering is chosen as 20% of the interquartile ranges.

First, we derive the overall score of concordant pairs for the sample $\{(1,1),\ldots,(n,n)\}$ if no outlier is present. Recall that if the paramter is automatically chosen as 20% of the interquartile ranges, we have $r_x = r_y = \frac{n}{10}$. In this setting, the i-th row $(1 \leq i \leq n-1)$ of both matrices consist of i zeros, followed by $\left\lfloor \frac{n-1}{10} \right\rfloor$ values between zero and one and $n - i - \left\lfloor \frac{n-1}{10} \right\rfloor$ ones. An exception is the last $\left\lfloor \frac{n-1}{10} \right\rfloor$ rows for which row i contains only $n-i$ values between zero and one. The values ranging from zero to one are given by

$$\frac{i}{\frac{n}{10}} = \frac{10i}{n} \text{ for } 1 \leq i \leq \left\lfloor \frac{n-1}{10} \right\rfloor .$$

In case that k outliers that result in the largest possible decrease of $\tilde{\gamma}$ are present, the values in the matrix M_X remain unchanged whereas for $M_{Y'}$, the first k rows contain only zeros and the first k columns contain only ones below the main diagonal. Hence, the values $M_{X,i,j}$ and $M_{Y',i,j}$ coincide if $i > k$ and $j > i$. In case of the minimum t-norm $T_{\mathbf{M}}$, it holds that $T_{\mathbf{M}}(t,t) = t$ for all $t \in (0,1)$ and consequently, the overall score of concordant pairs is

$$\tilde{C}_{n,k} = \sum_{i=k+1}^{n-1-\left\lfloor \frac{n-1}{10} \right\rfloor} \left(n - i - \left\lfloor \frac{n-1}{10} \right\rfloor \right) + (n-k-1) \sum_{i=1}^{\left\lfloor \frac{n-1}{10} \right\rfloor} \frac{10i}{n} - \sum_{i=1}^{\left\lfloor \frac{n-1}{10} \right\rfloor - 1} \sum_{j=1}^{i} \frac{10\left(\left\lfloor \frac{n-1}{10} \right\rfloor + 1 - j \right)}{n} .$$

Moreover, for the overall score of discordant pairs it holds that

$$\tilde{D}_{n,k} = \sum_{i=1}^{k} \left(n - 1 - \left\lfloor \frac{n-1}{10} \right\rfloor \right) + k \sum_{i=1}^{\left\lfloor \frac{n-1}{10} \right\rfloor} \frac{10i}{n} .$$

For large n, the expression $\tilde{\gamma}_{n,k} = \frac{\tilde{C}_{n,k} - \tilde{D}_{n,k}}{\tilde{C}_{n,k} + \tilde{D}_{n,k}}$ simplifies to[1]

$$\tilde{\gamma}_{n,k} = \frac{271n^3 + 600k^2 n - 343n - 60k(19n^2 - 9) + 72}{271n^3 - 343n + 72} .$$

[1] For large n, the floor function does not have any influence on the fraction.

Then, set $k = an$ (with $0 \leq a < 0.25$) and calculate $\lim_{n\to\infty} \tilde{\gamma}_{n,k}$ to obtain

$$\frac{600a^2 - 1140a + 271}{271} \,,$$

which, for $a \to 0.25^-$, yields $\frac{47}{542}$. The other fractional limits in Table 1 can be derived in a similar manner.

S2 Finite Sample Gross Error Sensitivity for Fixed Parameter

For $R_\varepsilon^{\text{crisp}}$, x_n needs to be chosen such that $x_n > (n-1) + \varepsilon$ holds, whereas for the R_r^{lin}, $x_n - (n-1)/r \geq 1$, i.e $x_n \geq r + n - 1$ needs to be fulfilled. For both R_b^{exp} and R_σ^{Gauss}, x_n needs to be as large as possible.

In all cases, the overall score of discordant pairs $\tilde{D}$ will tend to $n-1$, whereas $\tilde{C}$ will depend on the ordering, its parameter and the chosen t-norm. For $R_\varepsilon^{\text{crisp}}$, the choice of the t-norm is obviously irrelevant and

$$\tilde{C} = \begin{cases} \frac{1}{2}(n-2)(n-1) & \text{if } 0 < \varepsilon < 1, \\ \frac{1}{2}(n - \lfloor\varepsilon\rfloor - 2)(n - \lfloor\varepsilon\rfloor - 1) & \text{if } 1 \leq \varepsilon < n-1, \\ 0 & \text{if } n-1 \leq \varepsilon \,. \end{cases}$$

For R_r^{lin}, the choice of the t-norm influences the value of $\tilde{C}$. For $\bar{T} = T_{\mathbf{M}}$, we have

$$\tilde{C} = \begin{cases} \frac{(n-1)(n-2)}{2} & \text{if } r \leq 1, \\ \sum\limits_{i=1}^{g(r)} \frac{i(n-1-i)}{r} + \frac{(n-g(r)-2)(n-g(r)-1)}{2} & \text{if } 1 < r < n-1, \\ \frac{(n-1)\left((n-1)^2-1\right)}{6r} & \text{if } n-1 \leq r \,, \end{cases}$$

where

$$g(r) = \begin{cases} r-1 & \text{if } r \in \mathbb{Z}, \\ \lfloor r \rfloor & \text{if } r \notin \mathbb{Z} \,. \end{cases}$$

In case of the product t-norm $T_{\mathbf{P}}$, the value of $\tilde{C}$ is

$$\tilde{C} = \begin{cases} \frac{(n-1)(n-2)}{2} & \text{if } r \leq 1, \\ \sum\limits_{i=1}^{g(r)} \frac{i^2(n-1-i)}{r^2} + \frac{(n-g(r)-2)(n-g(r)-1)}{2} & \text{if } 1 < r < n-1, \\ \frac{n(n-2)(n-1)^2}{12r^2} & \text{if } n-1 \leq r \,. \end{cases}$$

Finally, for the Łukasiewicz t-norm $T_{\mathbf{L}}$, the value of $\tilde{C}$ is

$$\tilde{C} = \begin{cases} \frac{(n-1)(n-2)}{2} & \text{if } r \leq 1, \\ \sum\limits_{i=1}^{\lceil \frac{r-1}{2} \rceil} \left(n-1-\lceil \frac{r}{2} \rceil\right) \cdot \left(\frac{2\left(i + \lceil \frac{r}{2} \rceil - 1\right)}{r} - 1 \right) + \frac{1}{2}\left(n - \lceil r \rceil - 1\right)\left(n - \lceil r \rceil\right) & \text{if } 1 < r \leq n-2, \\ \sum\limits_{i=\lceil \frac{r}{2} \rceil}^{n-2} (n-1-i) \cdot \left(\frac{2i}{r} - 1\right) & \text{if } n-2 < r < 2(n-2) \\ 0 & \text{if } 2(n-2) \leq r \,. \end{cases}$$

S3 Supplementary Figures

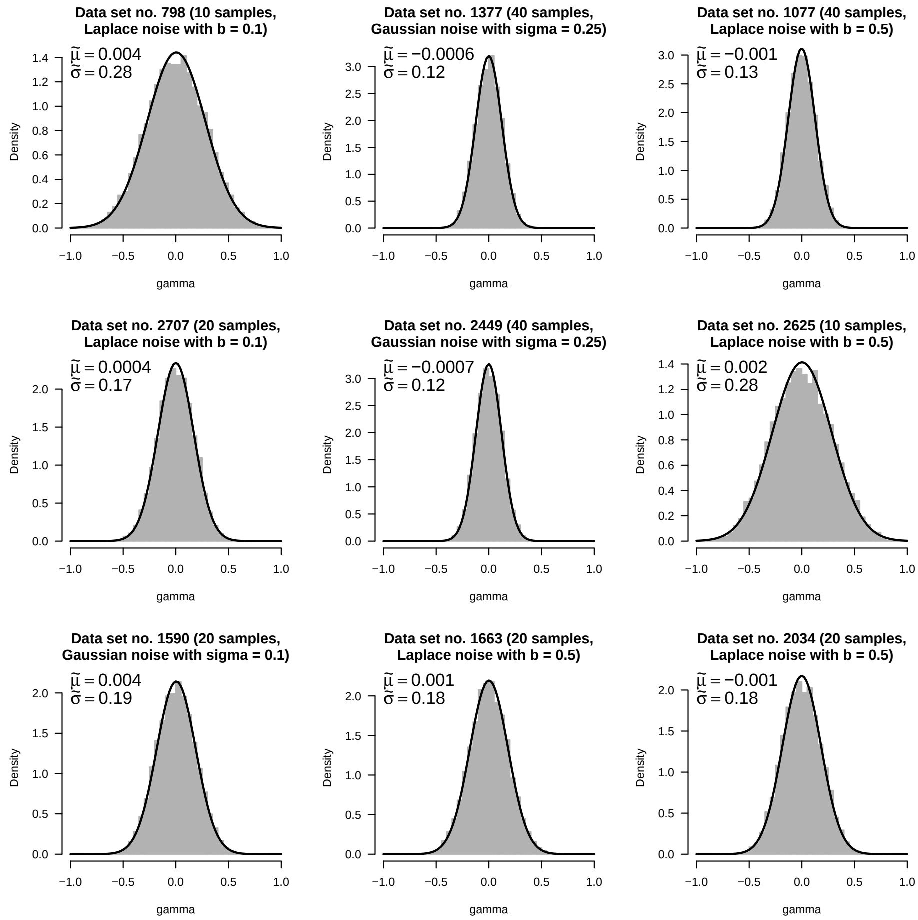

Figure S1: Histograms of gamma values of 100,000 random permutations of the 12 data sets shown in Figure 4 in the main paper. The gamma values were computed using the ad-hoc settings. The curves depict the distribution densities of the normal distributions with parameters estimated from the gamma values.

S4 Supplementary Tables

Table S1: Asymptotic gross error sensitivity for the combinations of ordering and t-norm in the `rococo` package if the parameter of the ordering is chosen as 20% of the interquartile ranges. Note that we only obtained upper bounds for some combinations of ordering and t-norm.

	$T_{\mathbf{M}}$	$T_{\mathbf{P}}$	$T_{\mathbf{L}}$
R_r^{lin}	4.428	4.589	4.674
R_b^{exp}	≤ 4.89	≤ 5.45	≤ 5.72
R_σ^{Gauss}	≤ 5.20	≤ 5.67	≤ 5.90
$R_\varepsilon^{\text{crisp}}$	4.938	4.938	4.938

Table S2: Pairwise comparisons of the p-values obtained for the seven compared methods on the test bed of 2,880 monotonically associated data sets. Each entry in this table gives the p-value of a one-sided paired Wilcoxon test applied to all 2,880 p-values of the two tests under consideration. So each entry gives the p-value of the test that the "row test" produces smaller p-values than the "column test".

	Spearman	Gaussian	Kendall	gamma	robust gamma (ad-hoc settings)	robust gamma (using prior inf.)	robust gamma (noise-tolerant)
Spearman	—	~ 1.0	~ 1.0	~ 1.0	~ 1.0	~ 1.0	~ 1.0
Gaussian	$2.0 \cdot 10^{-116}$	—	$1.6 \cdot 10^{-226}$	$3.2 \cdot 10^{-79}$	$3.4 \cdot 10^{-9}$	~ 1.0	$5.0 \cdot 10^{-13}$
Kendall	$1.9 \cdot 10^{-14}$	~ 1.0	—	~ 1.0	~ 1.0	~ 1.0	~ 1.0
gamma	$8.6 \cdot 10^{-9}$	~ 1.0	$1.5 \cdot 10^{-50}$	—	~ 1.0	~ 1.0	~ 1.0
robust gamma (ad-hoc settings)	$1.3 \cdot 10^{-76}$	~ 1.0	$1.7 \cdot 10^{-199}$	$2.9 \cdot 10^{-90}$	—	~ 1.0	$1.3 \cdot 10^{-4}$
robust gamma (using prior inf.)	$1.1 \cdot 10^{-111}$	$6.8 \cdot 10^{-7}$	$1.5 \cdot 10^{-241}$	$4.9 \cdot 10^{-109}$	$1.9 \cdot 10^{-34}$	—	$9.4 \cdot 10^{-91}$
robust gamma (noise-tolerant)	$5.8 \cdot 10^{-29}$	~ 1.0	$1.7 \cdot 10^{-66}$	$1.6 \cdot 10^{-18}$	~ 1.0	~ 1.0	—

Table S3: Pairwise comparisons of the p-values obtained for the eight compared methods on the test bed of 6,000 data sets of independent vectors. Each entry in this table gives the p-value of a one-sided unpaired Wilcoxon(-Mann-Whitney) test applied to all 6,000 p-values of the two tests under consideration. So each entry gives the p-value of the test that the "row test" tends to produce smaller p-values than the "column test".

	Spearman	Gaussian	Kendall	gamma	robust gamma (linear)	robust gamma (Laplace)	robust gamma (Gauss)	robust gamma (noise-tolerant)
Spearman	—	0.549	0.010	0.073	0.563	0.602	0.610	0.158
Gaussian	0.451	—	0.007	0.054	0.512	0.553	0.563	0.129
Kendall	0.990	0.993	—	0.767	0.993	0.995	0.996	0.911
gamma	0.927	0.946	0.233	—	0.945	0.956	0.959	0.684
robust gamma (linear)	0.437	0.488	0.007	0.055	—	0.539	0.550	0.122
robust gamma (Laplace)	0.398	0.447	0.005	0.044	0.461	—	0.511	0.104
robust gamma (Gauss)	0.390	0.437	0.004	0.041	0.450	0.489	—	0.098
robust gamma (noise-tolerant)	0.842	0.871	0.089	0.316	0.878	0.896	0.902	—