Performance Evaluation of NPI Methods with Copula for Bivariate Data

Hadeer A. Ghonem [1] Tahani Coolen-Maturi [2] Frank P. A. Coolen [3]

[1,2,3]Department of Mathematical Sciences, Durham University, Durham, UK

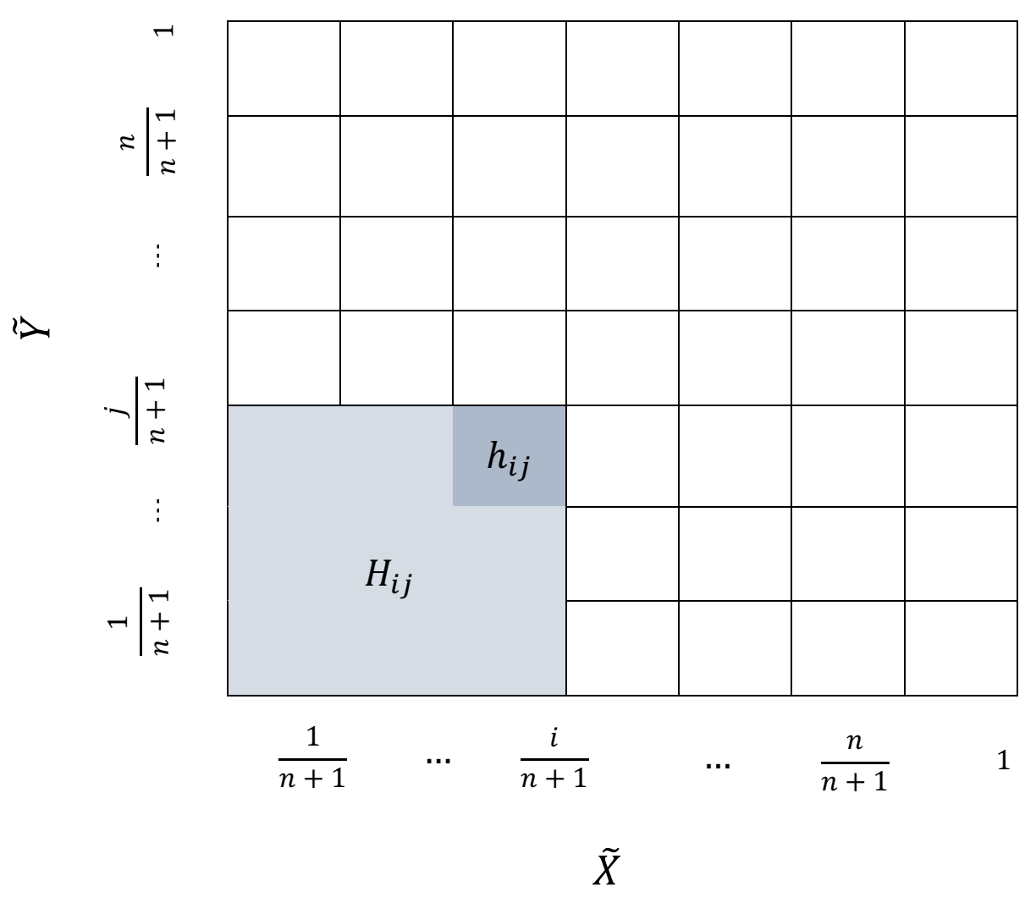
Abstract

The Nonparametric Predictive Inference (NPI) approach is based on Hill's assumption $A_{(n)}$ and uses imprecise probabilities to quantify uncertainty [1, 3]. It is interesting to assess the performance of NPI methods because of the imprecision involved. Furthermore, the existing methods used for evaluating the performance are mostly straightforward with precise probability but not trivial with imprecise probability. In this study the performance of the semi-parametric predictive method introduced by Coolen-Maturi et al. [2] has been evaluated in different aspects. A simulation study has been conducted to study the performance of this method using various measures, and there are two scenarios to consider. The first scenario assumes that the copulas used for simulating the data and performing the inference are the same, while the second scenario assumes that they are different. The coverage and width of the prediction intervals have been measured using different metrics. Moreover, the performance of this method has been investigated using loss functions and interval scores.

Introduction

Coolen-Maturi et al. [2] have introduced a semi-parametric predictive method which combines a parametric copula with NPI. This method uses NPI for the marginals and then a parametric copula is used to model the dependence between the variables and the parameter is estimated using the pseudo maximum likelihood method. Supposing n bivariate observations (x_i, y_i) where $i = 1, 2, ..., n$ observed from n bivariate random quantities. This approach involves predicting a future bivariate observation, denoted as (X_{n+1}, Y_{n+1}), using NPI for the marginals. Subsequently, these random quantities are transformed from the $[-\infty, \infty]^2$ plane to the $[0, 1]^2$ plane, resulting in $(\tilde{X}_{n+1}, \tilde{Y}_{n+1})$. Under the assumption $A_{(n)}$, the $[0, 1]^2$ plane is divided into $(n+1)^2$ squares of equal sizes, and the probability of $\tilde{X}_{n+1}$ falling in the interval $\left(\frac{i-1}{n+1}, \frac{i}{n+1}\right)$ is $\frac{1}{n+1}$, and similarly for $\tilde{Y}_{n+1}$ where $i, j = 1, 2, ..., n+1$.

The probability of the transformed future observation falling within any of the $(n+1)^2$ squares in the partitioned plane $[0, 1]^2$ is given by $h_{ij}(\hat{\theta}) = P_C\left(\tilde{X}_{n+1} \in \left(\frac{i-1}{n+1}, \frac{i}{n+1}\right), \tilde{Y}_{n+1} \in \left(\frac{j-1}{n+1}, \frac{j}{n+1}\right) |\hat{\theta}\right)$.



where $i, j = 1, 2, ..., n+1$, and $P_C(.|\hat{\theta})$ denotes the copula probability with the estimated parameter $\hat{\theta}$.

Example

Considering the event of interest $T_{n+1} = X_{n+1} + Y_{n+1}$, and the lower and upper probabilities for the event $T_{n+1} > t$ are considered. The lower probability is defined as the sum of $h_{ij}(\hat{\theta})$ for all $(i, j) \in L_t$, where $L_t = \{(i, j) : x_{i-1} + y_{j-1} > t, 1 \leq i \leq n+1, 1 \leq j \leq n+1\}$. The upper probability, on the other hand, is the sum of $h_{ij}(\hat{\theta})$ for all $(i, j) \in U_t$, where $U_t = \{(i, j) : x_i + y_j > t, 1 \leq i \leq n+1, 1 \leq j \leq n+1\}$.

A simulation study was conducted to assess the performance of the method, with the event of interest being $T_{n+1} > t$. The simulation of $n+1$ pairs of (x_i, y_i) where $i = 1, 2, ..., n+1$ was repeated N times and for each simulation run, the first n pairs are used as the data, and the last pair with $i = n+1$ is used as a future observation to assess the predictive inference. Considering the last pair of the run r $(r = 1, ..., N)$ is (x^r_{n+1}, y^r_{n+1}), then t^r_{n+1} is defined as $t^r_{n+1} = x^r_{n+1} + y^r_{n+1}$. For a given $q \in (0, 1)$, and $\underline{t}^r_q$ and $\overline{t}^r_q$ are the inverse values of the lower and upper probabilities for the event $T_{n+1} > t$, respectively. Here, l and u were defined as follows:

$$l = \frac{1}{N}\sum_{r=1}^{N} \mathbf{1}(t^r_{n+1} \geq \overline{t}^r_q) \qquad \text{and} \qquad u = \frac{1}{N}\sum_{r=1}^{N} \mathbf{1}(t^r_{n+1} \geq \underline{t}^r_q). \tag{1}$$

The aim of this performance evaluation method is to ensure that the inequality $l \leq q \leq u$ is satisfied.

Performance evaluation measures

The following measures were used to study the performance of the semi-parametric predictive method:

- Prediction Interval Coverage Probability (PICP)

$$PICP = \frac{1}{N_c}\sum_{k=1}^{N_c} \mathbb{I}_{[l^k,u^k]}(q), \text{ where } \mathbb{I}_{[l^k,u^k]}(q) \begin{cases} 1 & \text{if } q \in [l^k, u^k], \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

  Where N_c is the number of prediction intervals.
- Mean Prediction Interval Width (MPIW)

$$MPIW = \frac{1}{N_c}\sum_{k=1}^{N_c}(u^k - l^k). \tag{3}$$

- $MPIW_{q\in[l^k,u^k]}$ and $MPIW_{q\notin[l^k,u^k]}$: These measures separate intervals into two groups: those including q and those not including q. The average of width of intervals that include q, denoted as $MPIW_{q\in[l^k,u^k]}$, and the average width of the intervals that do not include q, denoted as $MPIW_{q\notin[l^k,u^k]}$.
- Quadratic Loss Function (L_Q) and Absolute Loss Function (L_{Abs}) Where (p) is the predicted value.

$$L_Q = (q-p)^2 \qquad \text{and} \qquad L_{Abs} = |q-p|. \tag{4}$$

  Where p is the predicted value.
- interval score

$$IS_{(c_1,c_2,c_3)}(l, u, q) = c_1(u-l) + c_2 \max\{0, l-q\} + c_3 \max\{0, q-u\}, \tag{5}$$

  where $\sum_{i=1}^{3} c_i = 1$ and $c_i \geq 0$.

Simulation study

In this study, the following algorithm is used to get N_c prediction intervals of $[l^k, u^k]$ then the previous performance evaluation measures can be used. Two scenarios are assumed here, the first scenario assumes the copula family used for simulating the data and inference is Normal. But the second scenario assumes Frank and Clayton copulas are used for inference and Normal copula for simulating the data.

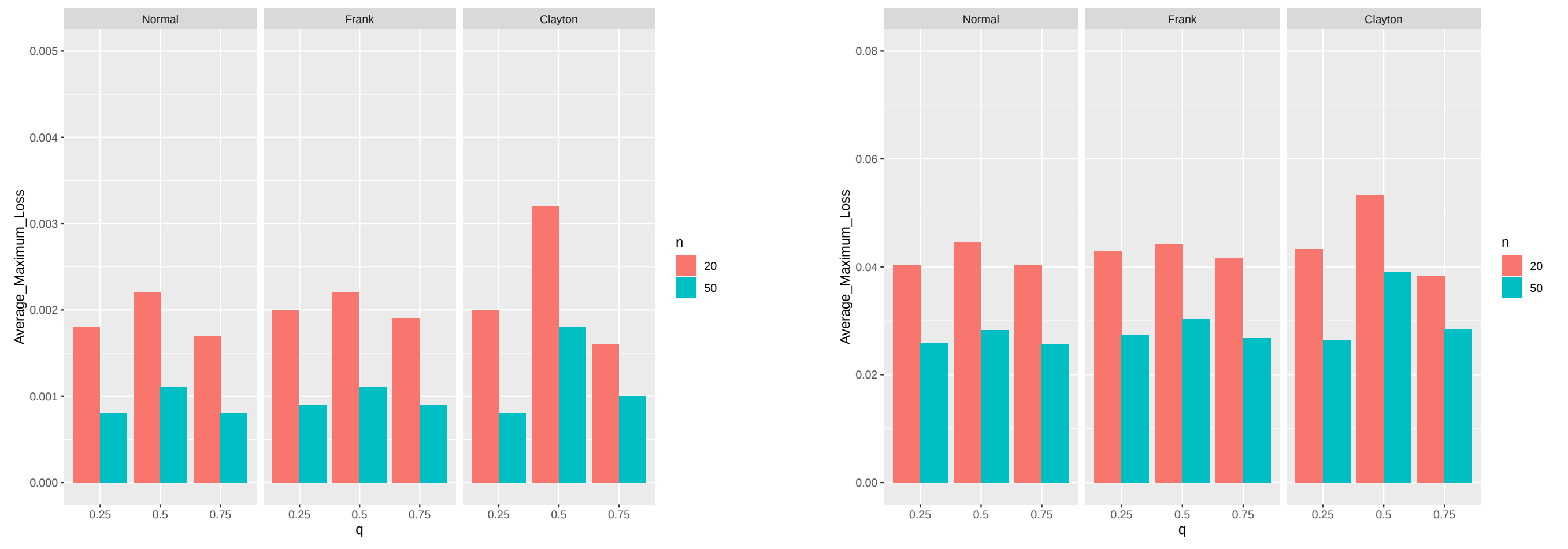
Algorithm

```
Algorithm 1: Computing performance evaluation measures
Result: Compute performance evaluation measures
for $k = 1$ to $N_c$ do
    for $r = 1$ to $N$ do
        Generate $n + 1$ pairs sample from a specified copula and use the first $n$ pairs for the semi-parametric method
          and the last pair is used for assessing the performance of the method;
        Estimate the copula parameter $\hat{\theta}$ using the first $n$ paired samples;
        Compute the probabilities $h_{ij}(\hat{\theta})$ ;
        Compute $t^r_{n+1}$=$x^r_{n+1}$+$y^r_{n+1}$ and $\underline{t}^r_q$ and $\overline{t}^r_q$;
        Compute $l$ and $u$ using;
    end
    Return $l^k$ and $u^k$;
end
```

Simulation results

Table 1. Simulation from Normal; when $q = 0.75$ and $\tau = 0.5$

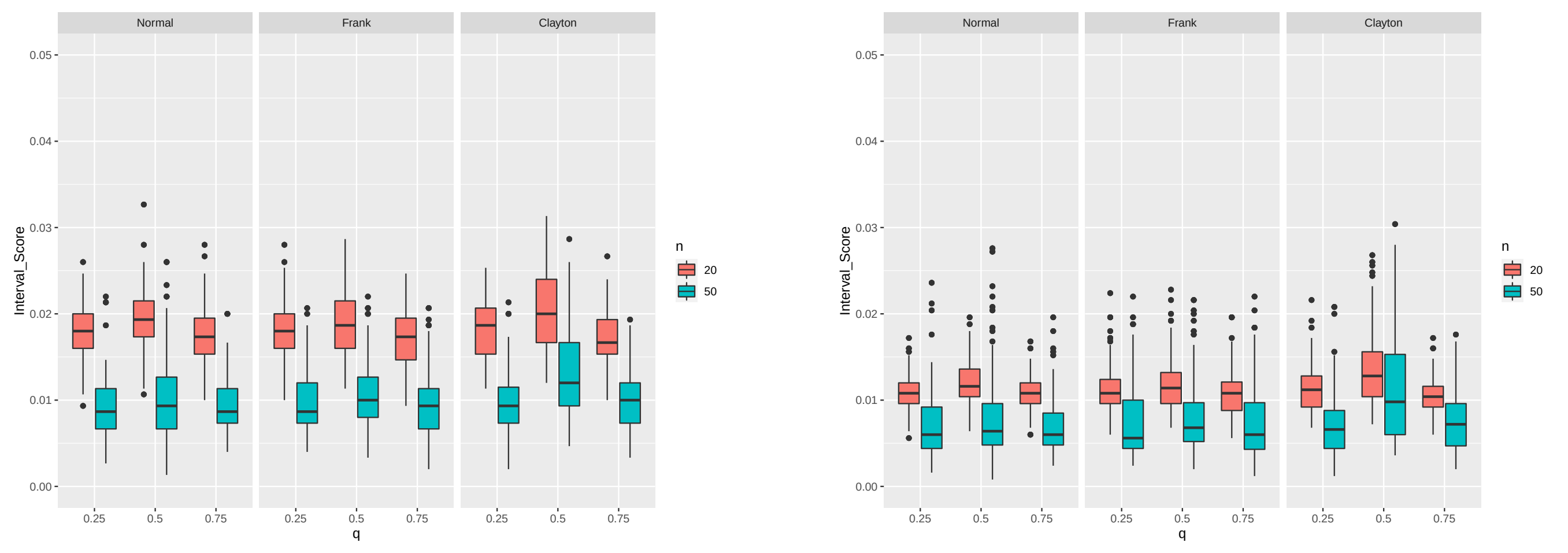
Normal copula is assumed for inference							
$n = 20$				$n = 50$			
$PICP$	$MPIW$	$MPIW_{q\in[l^k,u^k]}$	$MPIW_{q\notin[l^k,u^k]}$	$PICP$	$MPIW$	$MPIW_{q\in[l^k,u^k]}$	$MPIW_{q\notin[l^k,u^k]}$
0.87	0.0522	0.0530	0.0463	0.46	0.0219	0.0239	0.0202
Frank copula is assumed for inference							
$n = 20$				$n = 50$			
$PICP$	$MPIW$	$MPIW_{q\in[l^k,u^k]}$	$MPIW_{q\notin[l^k,u^k]}$	$PICP$	$MPIW$	$MPIW_{q\in[l^k,u^k]}$	$MPIW_{q\notin[l^k,u^k]}$
0.80	0.0497	0.0509	0.0450	0.43	0.0213	0.0219	0.0208
Clayton copula is assumed for inference							
$n = 20$				$n = 50$			
$PICP$	$MPIW$	$MPIW_{q\in[l^k,u^k]}$	$MPIW_{q\notin[l^k,u^k]}$	$PICP$	$MPIW$	$MPIW_{q\in[l^k,u^k]}$	$MPIW_{q\notin[l^k,u^k]}$
0.80	0.0492	0.0510	0.0421	0.34	0.0220	0.0229	0.0215



(a) Quadratic loss

(b) Absolute loss

Figure 1. Average of maximum losses when simulation from Normal and $\tau = 0.5$



(a) $c_1 = c_2 = c_3 = \frac{1}{3}$

(b) $c_1 = 0.2$, $c_2 = c_3 = 0.4$

Figure 2. Interval scores when simulation from Normal and $\tau = 0.5$

Discussion

The presented table 1 demonstrates a consistent pattern across all copula families used for inference, indicating that when the sample size is increased from 20 to 50, the values of performance evaluation metrics decreases. Notably, the comparing the values of $MPIW_{q\in[l^k,u^k]}$ and $MPIW_{q\notin[l^k,u^k]}$ reveals that intervals including the value q tend to be slightly wider than those that exclude q.

Figure 1 shows that whether employing a quadratic or absolute loss function, the average maximum loss of prediction intervals is higher for $n = 20$ compared to $n = 50$. This demonstrates that increasing the sample size leads to narrower intervals with the value q being close to the prediction intervals resulting in lower average maximum loss. Remarkably, when $q = 0.5$ with using Clayton copula for inference yields significantly larger average maximum loss compared to other scenarios. In addition, using Normal or Frank copulas for inference results prediction intervals with similar characteristics.

The results of interval scores are presented as box plots in Figure 2. It is observed that the interval scores tend to be higher for $n = 20$ compared to $n = 50$. In case of $c_1 = 0.2, c_2 = c_3 = 0.4$ there are slight differences between the interval scores for $n = 20$ and $n = 50$. When $n = 50$, smaller interval scores indicate better performance, even if the intervals do not include the value q in many cases.

Conclusion

The simulation results indicate that increasing the sample size leads to more true values falling outside the prediction intervals while the widths of the intervals decrease. However, the true values are close to the prediction intervals in case the intervals do not include the values. Using large sample sizes leads to smaller imprecision, which results in the prediction interval unlikely to include the real value. On this basis, the importance of using loss functions increases to measure the maximum and minimum distances between the real value and the intervals. Absolute loss function gives less consideration to the outliers than the quadratic loss function. Thus if the application pays attentions to outliers, quadratic loss function is advised to be used.

References

[1] T. Augustin and F.P.A. Coolen.
Nonparametric predictive inference and interval probability.
Journal of Statistical Planning and Inference, 124(2):251–272, 2004.

[2] T. Coolen-Maturi, F. P. A. Coolen, and N. Muhammad.
Predictive inference for bivariate data: Combining nonparametric predictive inference for marginals with an estimated copula.
Journal of Statistical Theory and Practice, 10(3):515–538, 2016.

[3] B. M. Hill.
Posterior distribution of percentiles: Bayes' theorem for sampling from a population.
Journal of the American Statistical Association, 63(322):677–691, 1968.