Performance Evaluation of NPI Methods with Copula for Bivariate Data*

Hadeer A. Ghonem HADEER.A.GHONEM@DURHAM.AC.UK
Tahani Coolen-Maturi TAHANI.MATURI@DURHAM.AC.UK
Frank P. A. Coolen FRANK.COOLEN@DURHAM.AC.UK
Department of Mathematical Sciences, Durham University, Durham, UK.

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 are mostly explicit with precise probability but not straightforward with imprecise probability. Therefore, the need to develop new measures to evaluate the performance of NPI methods arises.

Studying the dependence structure of variables is essential in many applications as it helps in understanding their relationships. This understanding enables inference based on a bivariate data set. Coolen-Maturi et al. [2] have introduced a semi-parametric predictive method that combines a parametric copula with NPI. This method uses NPI for the marginals and then estimates the parameter by assuming a bivariate parametric copula. It was evaluated using simulation studies [2]. However, the evaluation only focused on measuring one prediction interval per case and did not consider all aspects of its performance. Additionally, larger sample sizes may lead to less imprecision. Therefore, further investigations are necessary to assess the method's performance while considering the imprecision degree.

In this study the performance of the semi-parametric predictive method is evaluated by measuring the coverage and width of the prediction intervals using various metrics. The Prediction Interval Coverage Probability (PICP) and Mean Prediction Interval Width (MPIW) are the primary measures used, along with two additional measures to separate the prediction intervals of whether they include the true value or not. Moreover, it is interesting to investigate the performance of this method using loss functions and interval scores. Thus, two loss functions used here: quadratic loss function and absolute loss function. A simulation study was conducted to study the performance of this method using these 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 results of this study 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.

This study reveals that 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 as the real value is mostly close to the prediction interval.

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. doi:https://doi.org/10.1016/j.jspi.2003.07.003.

[2] Tahani Coolen-Maturi, Frank P. A. Coolen, and Noryanti 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. doi:https://doi.org/10.1080/15598608.2016.1184112.

[3] Bruce M. Hill. Posterior distribution of percentiles: Bayes' theorem for sampling from a population. *Journal of the American Statistical Association*, 63(322):677–691, 1968. doi:https://doi.org/10.2307/2284038.

*This work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) Doctoral Studentship at the University of Durham.