The Journal of Open Source Software

NEEP: null empirically estimated p-values for high-throughput molecular survival analysis

Sean West[1], **Hesham Ali**[1], **and Dario Ghersi**[1]

1 School of Interdisciplinary Informatics, University of Nebraska at Omaha

DOI: 10.21105/joss.02044

Software

- Review
- Repository
- Archive

Editor: Mark A. Jensen

Reviewers:

- @SiminaB
- @majensen

Submitted: 19 December 2019
Published: 31 August 2020

Summary

When conducting survival analysis for molecular expression, a researcher has two main options: a regression test using the continuous expression as the independent variable with Cox Proportional Hazards (CPH) or a logrank test after separating patients into two groups such as low- vs high-expression with the Kaplan-Meier (KM) method. Both methods depend on the proportional hazards assumption. When this assumption is violated, using a binary variable performs better. However, KM survival analysis requires that the patients are split into two groups by a molecular expression threshold. In many cases, such as molecular survival analysis, a single threshold choice cannot be justified based on experimental design. The choice of a threshold when splitting up a continuous variable has been shown to be sensitive to patient group re-sampling (Sehgal et al., 2015). And in practice, small changes in the chosen threshold produce proportionally larger differences in the set of significant logrank tests (West, Kumar, Batra, Ali, & Ghersi, 2019).

To circumvent the need for an ad hoc molecular expression threshold, the logrank test can be calculated across a range of thresholds, producing a range of p-values. Choosing the minimum p-value from this range identifies the optimal split of patients into two groups, given a single molecular expression vector. However, taking the lowest p-value from a range will produce a non-Uniform(0,1), right-skewed distribution of p-values. Since p-values should be uniform under the null distribution, the skewed distribution cannot be used for valid statistical analysis. An equation was developed that could predict the correct p-values (Lausen & Schumacher, 1992); however, the precision given by the original authors is not precise enough for p-value correction procedures which are sensitive to very small p-value changes, as in the case of molecular analyses having many tens of thousands of observations. Thus, we developed NEEP, which overcomes this issue by sampling the null distribution using a bootstrapping procedure, which is parallelized to improve execution time performance. (West et al., 2019).

The null distribution is constructed by repeatedly permuting the patient order. For each permutation, the minimum p-value is calculated and added to the null distribution. Separately, the minimum p-value is obtained for each molecular expression vector. This null distribution is used to empirically determine the true p-values for this list of molecular expression vector p-values. In this way, the precision of the procedure is dependent on the number of samples (permutations) used to generate the null distribution. Since the null distribution is generated from the same set of patients, the corrected p-values are guaranteed to be Uniform(0,1) under the null if random. In other words, the minimum p-values for a set of random molecular expression vectors is the same as the null distribution. Because of this, the type 1 error and the FDR are guaranteed to be controlled. Finally, NEEP conducts False Discovery Rate (FDR) p-value correction (Benjamini & Hochberg, 1995) and calculates effect sizes, the hazard ratio and the 1, 2, and 5 year mortality ratios.

Statement of Need

Research purpose: `NEEP` offers non-parametric, high-throughput, and statistically valid survival analysis of molecular expression vectors.

Problem solved: In molecular expression, the choice of a single threshold to separate patients into low- and high-expression cannot be justified and produces variable results. `NEEP` chooses the threshold which maximizes the test statistic relating molecular expression and patient survival for each molecular expression vector. Then `NEEP` corrects the resulting biased p-value distribution using a empirically determined null distribution. To generate the null, `NEEP` permutes different orders of patients and calculates their p-values across a range in order to produce a `Null Empirically Estimated P-value` for each molecular expression vector. `NEEP` produces p-values which are uniform under their null distribution so that their precision is dependent on the size of the null generated. By doing so, `NEEP` circumvents the issue of choosing a single threshold while addressing the issue of asymmetric p-value when optimizing the relationship between molecular expression and patient survival.

Target audience: The target audience is anyone conducting molecular, high-throughput survival analysis that does not have confounding clinical variables and whose expression vectors may violate CPH assumptions. Full documentation is available in the project repository.

Acknowledgements

This project was partly funded by a University of Nebraska Collaboration Initiative/ System Science Seed Grant to Sushil Kumar, Hesham Ali, and Dario Ghersi and by the NIH AA026428 R21 grant to Sushil Kumar. The funder website is https://nebraska.edu/collaboration-initiative. The funders had no role in project design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, *57*(1), 289–300. doi:10.1111/j.2517-6161.1995.tb02031.x

Lausen, B., & Schumacher, M. (1992). Maximally selected rank statistics. *Biometrics*, 73–85. doi:10.2307/2532740

Sehgal, V., Seviour, E. G., Moss, T. J., Mills, G. B., Azencott, R., & Ram, P. T. (2015). Robust selection algorithm (rsa) for multi-omic biomarker discovery; integration with functional network analysis to identify miRNA regulated pathways in multiple cancers. *PloS one*, *10*(10), e0140072. doi:10.1371/journal.pone.0140072

West, S., Kumar, S., Batra, S. K., Ali, H., & Ghersi, D. (2019). Uncovering and characterizing splice variants associated with survival in lung cancer patients. *PLoS Comput Biol*, 15(10): e1007469. doi:10.1371/journal.pcbi.1007469