A weighted form of the estimator for the cause-specific cumulative incidence function can be computationally more efficient using R

Alessandro Gasparini

Introduction

The standard estimator for the cumulative incidence function can be written in many alternative forms [1]: for instance, it can be written as a weighted empirical cumulative distribution function or as a product-limit estimator.
As a consequence, it is possible to analyze competing risks data using weighted versions of standard survival analysis methods. Any software that allows the use of time-dependent weights in survival analysis methods can be used for estimation.
The aim of this study consists in exploring differences in terms of computational burden between ordinary and weighted methods. We will focus on the software R, as methods to estimate weights, and time procedures are readily available.

Methods

First, we generated competing risks data with different sample sizes (n = 100 to 10,000 individuals). We hypothesized possible failure from two competing events, one of which is considered to be of interest.
We generated values for age (μ = 50 years, SD = 10) and for an hypothetical treatment (factor with two levels, "treated" and "non-treated") to use as covariates. Failure times for each of the two events (plus censoring) were generated from exponentially distributed random variables with pre-specified failure rate.

Consequently, we performed two different simulations.

The first one aimed to compare the time necessary to estimate an ordinary Fine & Gray model (`crr` function from the `cmprsk` package) versus the time necessary to perform the weighted procedure. The weighted procedure required two steps: first, estimate weights using the `crprep` function from the `mstate` package; then, estimate a weighted Cox model using the `coxph` function from the `survival` package.
The second simulation hypothesized the need to estimate a competing risks model m times (with $m > 1$). Therefore, we compared the time necessary to run `crr` m times versus the time necessary to estimate weights once and run `coxph` m times. Both models in both simulations included age and treatment as covariates.

Each simulation was benchmarked 1,000 times using the `microbenchmark` package on a 8-threaded CPU system, without any explicit parallelization.

R code is available on Github [2].

Results

The first simulation showed that the weighted procedure requires more time to estimate a single Fine & Gray model compared to the ordinary estimator [Figure 1]. The advantage was consistent across sample sizes, but its magnitude varied from 1.2-fold to 9-fold, comparing the ratio of median times. Mann-Whitney-Wilcoxon U-Test comparing the time required by the two procedures was significant at every sample size (p-value < 0.01).

The second simulation [Figure 2], on the contrary, showed that for small n and m (n = 100 or 1,000, m = 10, respectively) the ordinary procedure is still somewhat quicker. Increasing m and n showed a growing advantage of the weighted procedure compared to the ordinary one: the ratio of median times decreased from around 1.1 for n = 100 and m = 10 to around 0.2 for n = 10,000 and m = 100, corresponding to a 80% decrease in time if using the weighted procedure. As in the previous simulation, U-Tests are significant (p-value < 0.01) at every combination of sample size n and number of models m.

References

1. *R. B. Geskus*, "Cause-specific cumulative incidence estimation and the Fine and Gray model under both left truncation and right censoring", Biometrics, 67, 39–49 (2011)

2. *https://github.com/ellessenne/crr-vs-crprep*

Figure 1: Results of Simulation 1

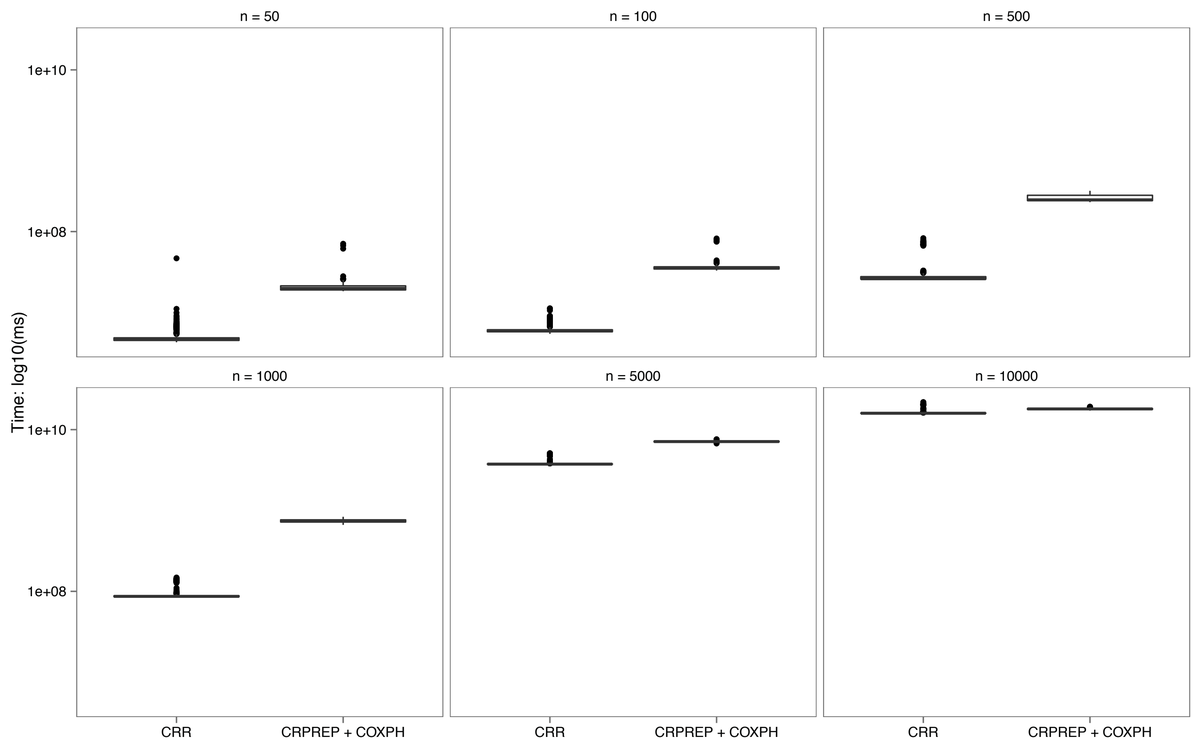
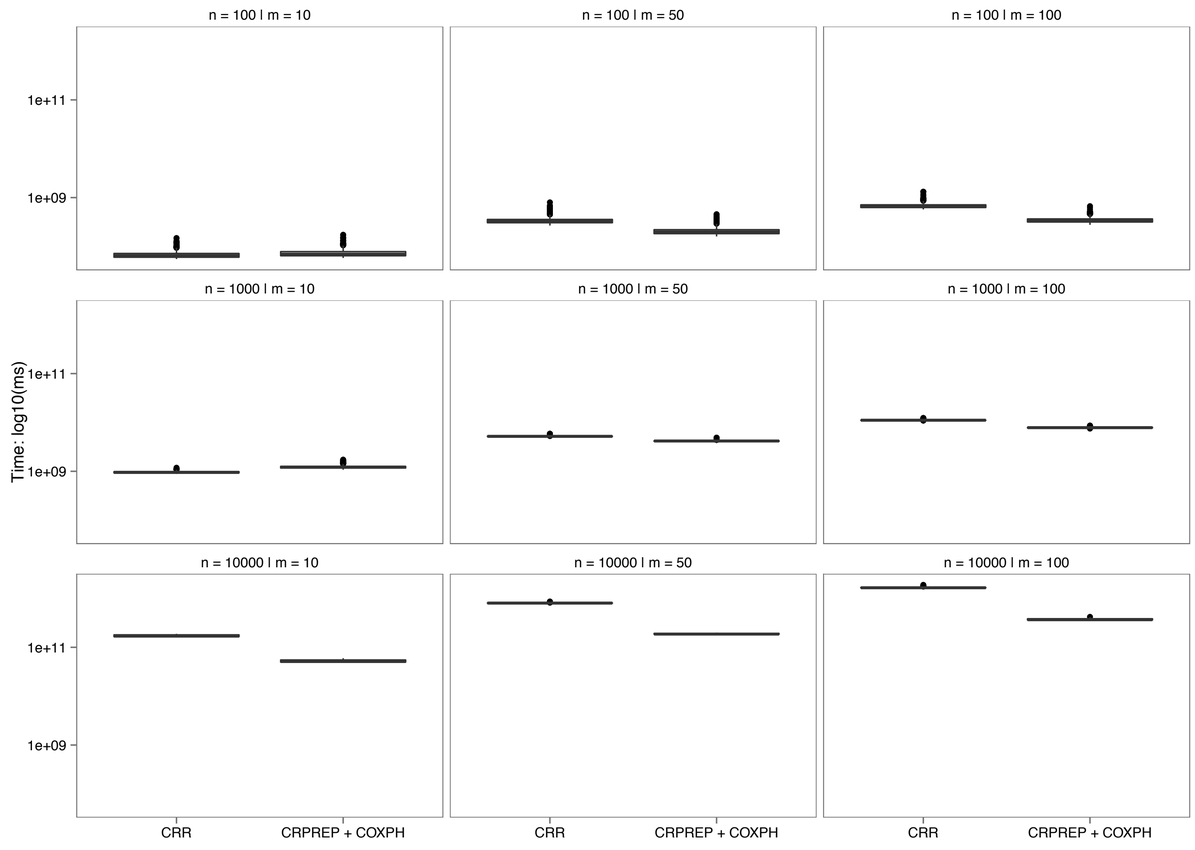


Figure 2: Results of Simulation 2



Conclusions

Geskus [1] first developed the alternative form of the standard cause-specific cumulative incidence function estimator in order to account for left truncation and right censoring when analyzing competing risks data. An interesting side property of that procedure is that it allows the use of weighted version of standard survival analysis methods, and those methods are often more computationally efficient compared to standard competing risks ones.
We showed that for analyses involving the estimation of a small number of models, the regular estimator and the `crr` function are to be preferred, although the advantage gets smaller as sample size increases.
Anyway, the bottleneck of the weighted procedure appears to be the estimation of weights. Therefore, in a situation where the estimation of weights needs to be done just once and is followed by the estimation of many models (such as multiple imputation of missing data, model selection, …), the weighted procedure can be more efficient in terms of computational time. Furthermore, such advantage of the weighted procedure increases steadily with sample size and number of models to estimate.

Alessandro Gasparini
a.gasparini9@campus.unimib.it
Renal Medicine and Baxter Novum (CLINTEC) – Karolinska Institutet, Stockholm, Sweden
Department of Statistics and Quantitative Methods, Università degli Studi di Milano-Bicocca, Milan, Italy

Karolinska Institutet