Master thesis proposal
Simulating from contraction type distributions

Krzysztof Bartoszek

October 19, 2017

Background — Contraction type distributions

Branching–processes have found applications in most contemporary branches of science. They provide stochastic models for phenomena that may be described by a tree structure. From computer science this can be the number of comparisons in the `Quicksort` algorithm. From evolutionary biology it can be the total path length of a tree—the sum of all root to node distances. In fact these two are described by the same probabilistic model. One is usually interested in the long–time growth (as the number of objects to sort increases) of such phenomena. Often it turns out that in the limit (as the number of objects becomes infinite) after proper scaling and centring the distribution F can be described as follows

$$Y \stackrel{\mathcal{D}}{=} \tau^r Y' + (1-\tau)^r Y'' + C(\tau), \tag{1}$$

where $Y, Y', Y'' \sim F$, $Y' \perp Y''$, $\tau \sim \text{Unif}[0,1]$ and $C(\tau)$ is some function. In fact for the limit of the number of comparisons of `Quicksort` we have $r = 1$ [3]. However, despite the apparent simplicity of Eq. (1) only rather recently was an exact EM algorithm proposed to simulate from the limit of `Quicksort`'s distribution [2]. No general method is known and other setups (e.g. $r \neq 1$) have to be worked with on a case by case basis. In [1] a heuristic approach (for $r = 2$ there) was considered that took advantage of the recursive form of Eq. (1).

Thesis project

The project is to implement and study heuristic methods to simulate from contraction type distributions. In some special cases we have an alternative description of the phenomena behind the distribution and are able to assess how well the implemented methods fare. This will give indications of how useful we expect the heuristics to be in other cases and also say something about their speed of convergence. In the scope of the work only the simplest model will be considered—the number of comparisons in the `Quicksort` algorithm (equivalently the total path length of a tree). Here direct method of simulations are available, by e.g. simulating a pure birth tree and then calculating its total path length.

Goals

The below general goals are for an "ideal" thesis. Depending on the student they will be made more specific in the direction of the student's interests. In particular the focus of the work will not be on the mathematical models (these will be "provided") but on implementing and putting together software to do simulations, inference and explore the statistical aspects of the models.

1. Become acquainted with basic graph theory for working with trees.
2. Become acquainted with pure birth branching process models and how to simulate them.
3. Become acquainted with contraction–type distributions.
4. Create a flexible R package that will allow for different methods of simulating from contraction–type distributions.
5. Study how do the results of the methods compare with the true distribution. Explore the basics of the theory behind comparing distributions.

Data

The topic will be illustrated with simulated data.

References

[1] K. Bartoszek. Exact and approximate limit behaviour of the Yule tree's cophenetic index. *ArXiv e-prints*, 2017.

[2] L. Devroye, J. A. Fill, and R. Neininger. Perfect simulation from the Quicksort limit distribution. *Electronic Comm. Probab.*, 5(12):95–99, 2000.

[3] U. Rösler. A limit theorem for "Quicksort". *Theor. Inf. Applic.*, 25(1): 85–100, 1991.