[PhD]
Statistical Analysis: Towards an Intuitive Code Editor

Duration : 3 years PhD grant (CIFRE)
Team : **Loki** (Inria Lille – Nord Europe & CRIStAL) – http://loki.lille.inria.fr/
Company : Zébrys – https://rplusplus.com
Advisor(s) : Stéphane Huot & Mathieu Nancel (first.last@inria.fr), C. Genolini (cg@rplusplus.com)

Data analysis is a complex task that requires writing computer code. Indeed, in many sensitive areas, analysis errors can have severe consequences such as wrong estimation of yet significant side effects of a drug. It is therefore critical to be able to verify the statistical analysis. This implies the ability to re-read code. But reading and writing code is a difficult task. And since statisticians are not always experienced programmers, it can be difficult for them to produce complex, robust, and bug-free code.

The objective of this thesis is to create an intuitive code editor, dedicated to statistical analysis and machine learning. It will have to simplify coding and to ease debugging.

Description of the thesis

Many users of computer tools for statistical analysis are occasional users for whom statistics are necessary but not their core expertise (medical doctors, psychologists, salesmen ...). They use various software (R, SAS, SPSS, Stata) that are generally poorly adapted to their needs and their expertise.

Common platforms for statistical analysis such as JMP [1] often provide users with graphical user interfaces to interactively manipulate and explore data (e.g. tables, interactive graph construction tools) or to build the analysis program without "code", using dialog boxes to choose between ready-to-use analysis modules. These "all integrated and interactive" approaches certainly reduce the entry cost and the learning efforts for a non-expert user in statistics and programming, and help to avoid bugs since they do not require to write code.

But these approaches reduce the control over possible analyzes and their parameters. For example, it is impossible to verify that there was no error in the computation of a complex statistic. Also, by promoting reuse of stereotyped analyzes, they give little guarantee on the relevance of chosen analyzes with respect to the problem studied, and probably do not allow to acquire more advanced skills in the field.

R++, the Next Step is a high performance statistical analysis software. It is integrated in a simple and user-friendly interface adapted to the expertise of users. Video presentations of the software ("R ++ in one minute") are available on our youtube channel: https://www.youtube.com/c/rplusplus. However, we now need to study and design interactive tools for better programming support, which is the goal of this thesis.

More appropriate approaches for exploring data and learning statistics have been proposed, such as Statsplorer [2], but they do not offer the power of a dedicated programming language like R does. There are also many advanced tools for supporting programming in general, integrated into IDEs (Integrated Development Environments): syntax highlighting, "smart" indentation and completion, real-time syntax checking, visual languages for program modeling, graphic and real-time debuggers, integrated help and tutorials, etc. Although they are generally relevant for writing and verifying statistical code, these generic tools can be further improved to take into account the specific needs of users in this area. In particular, it would be interesting, for example, to further explore dynamic visualizations of the links between the variables of the program, the analyzed data and the results of these analyzes. The guiding principles promoted by Bret Victor in his essays *Explorable Explanations* [3] [4] or *Learnable Programming* [5] are for example trails to explore in order to inform the design of such new tools.

Workplan for the thesis

The objective of the thesis is thus to study and implement these new interactive tools to simplify the production of the code, and to facilitate its proofreading and debugging in the context of statistical analysis. The methodology will consist of:

- Reviewing and studying state-of-the-art tools for data analysis and programming for statistical analysis, as well as for programming in general;
- Conducting a user-centered design approach (needs assessment, participatory design of coding tools adapted to statisticians, production and management of usage scenarios);
- Prototyping and evaluation of novel visualization and interaction techniques that meet these needs;
- Integration of these innovations in the code editor of R++, in collaboration with the engineers at Zébrys.

Références

[1] SAS, “JMP web page,” [Online]. Available: https://www.jmp.com/.

[2] C. Wacharamanotham, K. Subramanian, S. T. Völkel and J. Borchers, “Statsplorer: Guiding Novices in Statistical Analysis,” in *CHI '15 Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2015.

[3] B. Victor, “Explorable Explanations,” 2011. [Online]. Av. : http://worrydream.com/ExplorableExplanations/.

[4] B. Victor, “Up and Down the Ladder of Abstraction: A Systematic Approach to Interactive Visualization,” 2011. [Online]. Available: http://worrydream.com/LadderOfAbstraction/.

[5] B. Victor, “Learnable Programming: Designing a programming system for understanding programs,” 2012. [Online]. Available: http://worrydream.com/LearnableProgramming/.

Candidates

- The candidate must hold a Master degree in Computer Science or equivalent, with strong technical skills in programming;
- Basic knowledge in Human-Computer Interaction (participatory design, prototyping and evaluation) is a plus;
- Knowledge of statistics is a plus but not mandatory (expect to follow a bridging course).

Environnement

The thesis will take place in collaboration between the start'up Zébrys and the **Loki** research group at Inria.

- Zébrys is a start'up whose goal is to design the software *R ++, the Next Step*.
  R ++, the Next Step is a very high performance statistical analysis software. Up to 800 times faster than competitors, it natively integrates parallelism. It allows the exploitation of big data databases. And since not all statisticians are computer experts, it has been integrated into a modern and user-friendly graphical user interface.
  Zébrys is a start'up with high potential (French Tech laureate, CREALIA, Innov & Plus, Réseau Entreprendre)

- **Loki** is a research team in Human-Computer Interaction of the Inria Lille - Nord Europe research center, jointly with the CRIStAL laboratory (CNRS and University of Lille). Made up of internationally recognized experts in the field, the main goal of the team is to revisit the design of interactive systems to better take into account the capabilities and needs of their users, but also their designers. Among the notable expertise of the team, we can mention the study of human factors related to interaction, skills acquisition with digital tools, engineering of interactive systems.

Application

Send resume and cover letter to <cg@rplusplus.com> and <stephane.huot@inria.fr>