R++, the Next Step : une nouvelle interface graphique pour R

C. Genolini [a,b]

[a] Université Paris Nanterre
[b] Zebrys
cg@rplusplus.com

Mots clefs : Statistique, Interface Homme Machine, GUI, R++

Le projet

R++, the Next Step est un projet de développement d'une nouvelle implémentation de **R**. Il a pour vocation d'être compilable, d'intégrer en natif la gestion du parallélisme et de permettre l'exploitation des bases de données de grande dimension. Mais surtout, ***R++, the Next Step*** est intégré dans une interface homme machine moderne et conviviale, spécifiquement conçue pour simplifier l'analyse statistique.

L'Interaction Homme Machine

L'Interaction Homme Machine est la science ayant pour objectif d'étudier la manière dont les humains interagissent avec les ordinateurs afin d'ensuite concevoir des outils plus ergonomiques. Pour cela, des séances de brainstorming réunissant utilisateurs (dans notre cas les statisticiens) et informaticiens sont organisées. Dans un premier temps, l'objectif est de définir les tâches qui sont particulièrement ardues, pénibles à réaliser ou a fort risque d'erreur : tout ce qui fait cauchemarder les statisticiens. Ensuite, des solutions sont collectivement imaginées. Enfin, un prototype vidéo, illustration par l'exemple du problème et de sa solution, est élaboré.

Les « cauchemars » en statistique

Lors de séances de prototypage vidéo, nous avons identifiés différents aspects de l'analyse statistique particulièrement problématiques :

- **Ouverture des fichiers :** les fichiers que nous analysons proviennent de nombreuses sources, leur lecture n'est pas toujours simple (encodage, séparateur de colonne, lignes d'introduction dans un fichier,…)
- **Détection des valeurs aberrantes :** dans toutes les études, on trouve des étudiants qui ont 180 ans, ou qui mesurent 180m... La vérification doit se faire colonne par colonne. En outre, la position exacte d'une valeur aberrante peut être compliquée à déterminer, il est donc difficile de la corriger.
- **Le typage erroné :** une colonne d'entier qui contient la valeur 2O (deux puis la lettre O) sera identifié comme une colonne factor. Elle sera donc considérée comme un factor dans un summary ou dans une régression linéaire.
- **Fusion de modalités :** Homme, HOMME, homme et H seront considérées comme des modalités différentes. Identifier toutes les différentes versions d'une modalité puis les fusionner peut prendre du temps.
- **Ordinal ou factor :** Les variables ordonnées sont systématiquement considérées comme des factors. Il faut corriger à la main.

- **Modifier les graphes :** il est souvent nécessaire de passer par un éditeur externe comme The Gimp pour ajouter des détails sur les graphes.
- **Export des graphes :** l'export est également une opération délicate, en particulier quand l'éditeur nous demande de modifier les DPI ou le canal alpha…

Compilation des solutions

Pendant trois ans, en collaboration avec des expert IHM et des utilisateurs, nous avons cherché des solutions aux problèmes qui viennent d'être cités. Elles ont été présentées à des panels qui ont validé ce qui paraissait le plus adapté. Puis elles ont été compilées dans une interface unique.
Dans cet exposé, nous vous présenterons *R++*, *the Next Step*, une nouvelle interface graphique pour R qui permet de grandement simplifier l'analyse statistique.