Atouts et faiblesses du logiciel R en enseignement, recherche et industrie

Pierre-André Cornillon, Eric Matzner-Løber

Atouts et faiblesses de R en enseignement, recherche et industrie

Pierre-André Cornillon[1], & Eric Matzner-Løber[2]

(1) Montpellier SupAgro, pierre-andre.cornillon@supagro.inra.fr
(2) Université de Rennes, eml@uhb.fr

Résumé Dans cette conférence, nous aborderons le logiciel R sous trois aspects : son utilisation en enseignement, en recherche et dans le monde de l'entreprise. Pour chacune de ces trois utilisations, il est nécessaire d'évaluer les demandes spécifiques du domaine et les réponses qu'apporte R.

Rappelons brièvement que R est multi-plateforme et multi-OS. Il est entièrement gratuit, très complet et offre à la fois des commandes mais aussi des menus déroulants. Il est donc raisonnable de penser que ce logiciel fera partie des logiciels de statistique les plus enseignés.

Sa facilité de programmation et sa forte utilisation dans le monde de la recherche en font dès aujourd'hui un langage omniprésent. On peut donc s'interroger sur les futures évolutions de R vis à vis de la recherche en statistiques.

Dans la troisième partie, nous comparerons R avec ses différents concurrents et analyserons les points qui gouvernent les choix de logiciels en entreprise (prix, interface graphique, intégration dans les base de données...). Bien évidemment le logiciel est loin d'être parfait mais il comporte dès à présent des avantages qui lui valent d'être adopté par un nombre croissant d'entreprises.

Abstract

In this talk, we will present the use of R software in teaching, research and enterprise. To each of these three fields corresponds specifics problems.

Recall that R is truly multi-platform and multi-OS. This is a free software with numerous statistical methods already available. It offers also a Graphical User Interface in addition to the Command Line Interface, which is the natural way to interact with R. According to these remarks, it seems natural to think that R will be very soon the most used statistical software for teaching.

Its programming ease and its widespread use make R to be the lingua franca of the statistical community. What will be the future of R in statistical research ?

In the third part, we will try to make a comparison between R and other statistical softwares. We will analyze both positive and negative aspects of R software in the industry and try to understand if the success of R in research field will carry on in the industry.

Mots clés

R, package, logiciel libre, logiciel de statistique

Keywords

R, package, free software, statistical software

1 R et l'enseignement

Les méthodes statistiques sont toujours illustrées à partir d'exemples et donc de données. Afin de rendre les étudiants autonomes, l'enseignant doit donc prendre en compte le traitement de données dans la formation en statistiques. Les méthodes les plus simples, comme les calculs de moyennes ou de variances empiriques, peuvent être directement illustrées en utilisant des machines à calculer. Cependant, ce stade est rapidement dépassé et lorsque la méthodologie devient complexe, un logiciel doit être utilisé. A ce stade, l'enseignant se trouve confronté à plusieurs choix:

- utiliser un tableur et ses fonctionnalités ;
- utiliser un logiciel basé sur des commandes ou CLI (Command Line Interface), permettant d'enchaîner des commandes au sein scripts ;
- utiliser un logiciel basé sur des menus déroulant ou GUI (Graphical User Interface).

Au cours de cette présentation, nous envisagerons les avantages et les inconvénients de ces 3 approches et nous placerons le logiciel R dans ces trois possibilités.

Les avantages de R peuvent être résumés simplement. Il s'agit d'un logiciel gratuit qui est disponible sur de nombreux sites miroirs comme par exemple le site lyonnais: `http://cran.univ-lyon1.fr/`. Par ailleurs il est multi-plateformes (32 bits, 64 bits, Risc, Sisc) et multi-OS (windows, Mac, Linux, Unixes), ce qui facilite son intégration dans tous les environnements pédagogiques. Ensuite il s'agit d'un logiciel complet où presque toutes les méthodes statistiques sont déjà implémentées via des ajouts de librairies de programmes (ou packages) gratuits. Enfin, ce logiciel offre une base de type CLI avec des éditeurs adaptés (tinn-R, ESS pour emacs, kate etc.) qui permettent une intégration assez avancée de R. Cependant, une possibilité pédagogique originale de GUI est disponible via le package Rcmdr. Le logiciel R offre donc le choix à l'enseignant entre CLI et GUI sans changer d'environnement.

2 R et la recherche

Il semblerait que le langage proposé par S et R se rapproche du statut de langage commun de la communauté des statistiques. L'essor de ce langage peut être mesuré par la multiplication du nombre de packages disponibles pour R. Ainsi, le logiciel R constitue un des logiciels les plus complets des statistiques et surtout le plus réactifs aux nouvelles méthodes, comme le boosting, les svm ou le lasso.

Cet essor peut aussi se mesurer par le nombre en forte croissance des livres traitant de statistiques au travers du logiciel R. Des maisons d'édition comme Springer (par exemple [1]), Chapman & Hall (par exemple [7]) ou Wiley (par exemple [6]) proposent désormais de nombreux ouvrages sur ce logiciel en anglais et désormais des ouvrages en français

commencent à paraître [5]. Notons au passage que Springer possède même une collection spécifique intitulée useR!.

Enfin, jusqu'à présent les packages sont plutôt une conséquence d'un article dans une revue, comme par exemple le package mboost de Bühlmann & Hothorn [3] qui peut être vu comme une conséquence des articles antérieurs de Bühlmann sur le sujet [2,4]. Mais, afin d'assurer la reproductibilité des résultats numériques, la mise à disposition des codes et leur mise en forme semblent des objectifs raisonnables de l'évaluation scientifique des méthodes statistiques. L'élaboration de package R sera peut-être à l'avenir une nécessité dans la rédaction d'articles. Nous illustrerons la conception d'un package "basique" afin de monter la simplicité de la mise en œuvre.

3 R et le monde de l'entreprise

Lorsque l'entreprise utilise des méthodes statistiques elle est rapidement confrontée à des choix de logiciels. Même s'il est rare qu'une entreprise ne possède qu'un seul logiciel, les critères de choix sont relativement constants et relèvent des mêmes questionnements que nous pouvons résumer ainsi:

- Logiciel en production ou en investigation ?
- Interface graphique (GUI) ou non ?
- Intégration des logiciels vers les bases de données de l'entreprise ?
- Agrément FDA ?
- Logiciel payant et nécessité d'un interlocuteur ?
- Possibilité de programmer ses interfaces et de piloter des programmes dans d'autres langages de bas niveaux ?

Nous présenterons comment le logiciel R répond à ces questions et quelles sont ses avantages et inconvénients à l'heure actuelle. Même si des entreprises ont déjà fait le choix de R, ce logiciel ne répond pas actuellement à toutes les attentes.

4 Conclusion

Le logiciel R présente de nombreux atouts mais aussi des faiblesses. Il semble important de savoir, si celles-ci seront comblées. Rappelons que le projet R est un projet de type coopératif et il n'existe donc pas de prévisionnel concernant telle ou telle amélioration. Cependant, l'étendue des améliorations entre la première version de R et la version actuelle laisse à penser que de nombreuses améliorations seront réalisées et que ce logiciel constitue un choix d'avenir.

Bibliographie

[1] Albert, J. (2007). Bayesian computation with R. Springer, New-York.
[2] Bühlmann, P. (2006). Boosting for high-dimensional linear models. Annals of Statistics 34, 559-583.
[3] Bühlmann, P. & Hothorn, T. (2007). Boosting algorithms: regularization, prediction and model fitting (with discussion). Statistical Science 22, 477–505.
[4] Bühlmann, P. & Yu, B. (2003).Boosting with the l_2 loss: Regression and classification. Journal of American Statistical Association 98, 324-339.
[5] Cornillon, P-A., Guyader A., Husson F., Jégou N., Josse J., Kloareg M., Matzner-Løber E. & Rouvière L. (2008). Statistiques avec R, Presses Universitaires de Rennes, Rennes.
[6] Crawley, M. (2005). Statistics: An Introduction using R. Wiley, New-York.
[7] Faraway, J. (2007). Extending the linear model with R. Chapman & Hall/CRC, Boca Raton.