

Travailler avec des matrices creuses sous R

B. Thieurmel

DATAKnowledge
56 rue d'amsterdam -75009 Paris
bt@datak.fr

Mots clefs : Matrice creuse, Matrix, spam, irlba, slam, RcppEigen

Une matrice creuse peut se définir simplement comme une matrice contenant beaucoup de zéros. On peut par exemple se retrouver face à ces données dans les problématiques suivantes :

- Résolution d'équations aux dérivées partielles
- Traitement de données de composition (agro-alimentaires, cosmétiques, ...)
- Théorie des réseaux
- Système de recommandations (netflix)

Lorsque l'on est amené à travailler sur des matrices creuses, surtout si elles atteignent des tailles conséquentes, il peut vite être judicieux, voire obligatoire, d'utiliser un système et une structure de stockage appropriés, ainsi que des algorithmes adaptés.

Nous proposons ici un tour d'horizon rapide des outils disponibles sous R pour travailler sur les matrices creuses. Seront abordés :

1. Les possibilités de stockage, la plus simple et la plus intuitive étant par le biais d'un triplet (i, j, k) de vecteurs représentant respectivement les lignes, colonnes, et valeurs non-nulles de la matrice. Cependant, d'autres systèmes plus efficaces existent.
2. La manipulation des matrices creuses
3. Les algorithmes et opérations disponibles
4. Un exemple de codage de nouvelles fonctions, via l'utilisation de **RcppEigen**[4]

Nous nous intéresserons notamment aux packages **Matrix**[1], **spam**[2], **irlba**[3], et cela via une application sur des données réelles.

Références

- [1] Douglas Bates and Martin Maechler (2013). Matrix: Sparse and Dense Matrix Classes and Methods. R package version 1.0-12. <http://CRAN.R-project.org/package=Matrix>
- [2] Reinhard Furrer, Stephan R. Sain (2010). spam: A Sparse Matrix R Package with Emphasis on MCMC Methods for Gaussian Markov Random Fields. Journal of Statistical Software, 36(10), 1-25. <http://www.jstatsoft.org/v36/i10/>.
- [3] Jim Baglama and Lothar Reichel (2014). irlba: Fast partial SVD by implicitly-restarted Lanczos bidiagonalization. R package version 1.0.3. <http://CRAN.R-project.org/package=irlba>
- [4] Douglas Bates, Dirk Eddelbuettel (2013). Fast and Elegant Numerical Linear Algebra Using the RcppEigen Package. Journal of Statistical Software, 52(5), 1-24. <http://www.jstatsoft.org/v52/i05/>.