Sparse Model Matrices for Generalized Linear Models

Martin Mächler[1,2,*], **Douglas Bates**[1,2]

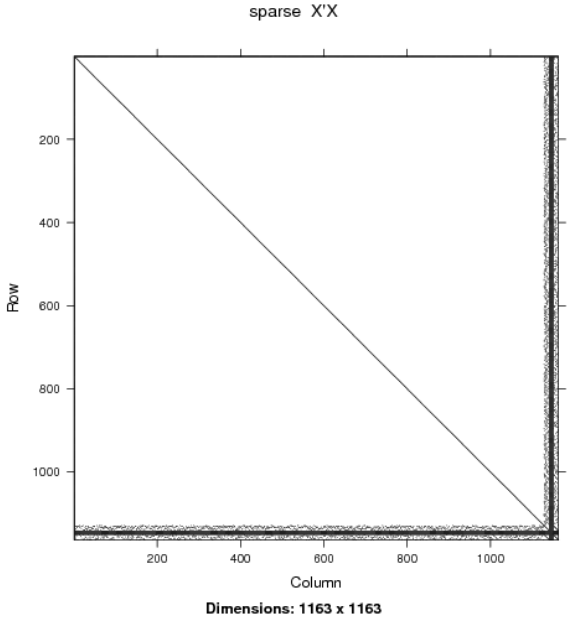
1. ETH Zurich, Switzerland and University of Wisconsin, Madison, USA
2. R Core Development Team *Contact author: maechler@stat.math.ethz.ch

Keywords: sparse matrices, linear models, GLM, mixed effects, large data

Using sparse model matrices for linear, generalized linear, and also (generalized) linear mixed effect models can be very advantageous particularly, when most predictor variables are categorical (`factor`s). We will demonstrate using such sparse matrices via function `sparse.model.matrix()` from the **Matrix** package and use sparse Cholesky decompositions to fit large linear and generalized linear models (GLM) efficiently. We will compare these with the corresponding functionality in R packages **biglm** and **speedglm**.

However, speed is not everything, and a not so well-known fact is that S and R have had smart "home-made" low-level code underlying `lm()`, for situations of ill-conditioned ("quasi-singular") design matrices $\mathbf{X}$. The benefit of this pivoting code has been the easy "identification" of components $\hat{\beta}_j$ to be set to `NA` such that other parts remained stably defined. Traditional high-quality code for matrix decompositions and least squares computation, as provided, e.g., by LAPACK, would not automatically provide the same functionality, nor do the analogous libraries for sparse computations.

We will explore some of these computational challenges in a mixed-effect model of moderately large data sets (e.g., $n = 73421$, $p = 1163$, $q = 2972$). Notably, the computation of standard errors for the fixed effects can pose problems which seem harder to solve than the parameter (maximum likelihood) estimation itself.



References

Douglas Bates and Martin Maechler (2005 ff). Introduction to the Matrix package (and other vignettes). `http://cran.r-project.org/web/packages/Matrix/`

Douglas Bates (2005 ff); Diverse vignettes on using Mixed-Effect models `http://lme4.r-forge.r-project.org/`

Douglas Bates (2010). **lme4: Mixed-Effects Modelling with R**; Springer, useR! series.

Thomas Lumley (2009). `biglm`: bounded memory linear and generalized linear models. R package version 0.7. `http://CRAN.R-project.org/package=biglm`

Martin Maechler and Douglas Bates (2009). Sparse Matrices in package Matrix and applications; slides from useR! 2009, Rennes. `http://matrix.r-forge.r-project.org/slides/2009-07-10-Rennes/MM-talk.pdf`