factorMerger: a set of tools to support results from post hoc testing

Agnieszka Sitko *University of Warsaw*

factorMerger is an *R* package whose purpose is to extend methods of analysing dependencies between groups of a categorical variable after carrying out an analysis of variance (ANOVA). The idea of the package arose from the need to create an algorithm which outputs in a hierachical interpretation of relations between levels of a categorical variable. Thereby, for a given significance level groups may be devided into nonoverlapping clusters. **factorMerger** implements iterative version of post hoc testing based on likelihood ratio test for parametric models: gaussian, binomial and survival. It also provides custom visualizations for each model built on **ggplot2** package.
Package webpage: https://github.com/geneticsMiNIng/FactorMerger

Keywords: analysis of variance (ANOVA), hierarchical clustering, likelihood ratio test (LRT), post hoc testing

Introduction

If data is analysed using ANOVA a more detailed analysis of differences among categorical variable levels might be needed. The traditional approach will be to perform *pairwise post hocs* - multiple comparisons after ANOVA. For each pair of groups we run specific statistical test which outputs with an information whether response averages in those groups are significantly different. However, if we look from the distant perspective, for a certain significance level, these results may not be consistent. One may consider the case that mean in group A does not differ significantly from the one in group B, similarly with groups B and C. In the same time difference between group A and C is detected. Then data partition is unequivocal and, as a consequence, impossible to put through.

The problem of clustering categorical variable into non-overlapping groups has already been present in statistics. First, J. Tukey proposed an iterative procedure of merging factor levels based on studentized range distribution (Tukey, 1949). However, statistical test used in this approach made it limited to gaussian models. Collapse And Shrinkage in ANOVA (CAS-ANOVA, Bondell and Reich (2008)) is an algorithm that extends categorical variable partitioning for generalized linear models. It is based on the Tibshirani's Fused LASSO (Tibshirani et al., 2005) with the constraint taken on the pairwise differences within a factor, which yields to their smoothing.

Delete or Merge Regressors algorithm (Prochenka, 2016*a*) is also adjusted to generalized linear models. It directly uses hierarchical clustering to gain hierarchical structure of a factor. At the beginning, *DMR4glm* calculates likelihood ratio test statistics for models arising from pairewise merging of factor levels against the initial model (the one with all groups included). Then performs agglomerative clustering taking LRT statistic as a distance — each step of clustering is associated with model with different factor structure. Experimental studies (Prochenka, 2016*b*) showed that Delete or Merge Regressors's performance is better than CAS-ANOVA's when it comes to model accuracy.

factorMerger package gives an approximate implementation of *DMR4glm*. In addition to the base algorithm, it also provides its sequential version, which merges only those levels which are

relatively close (levels distance is dependent on the model chosen). While the basic approach (all vs. all comparisons) may sometimes result in a slightly better partition from the statistical point of view, proposed extention (all vs. subsequent comparisons) seems to be more graceful when it comes to the interpretation. Moreover, the former is more computationally expensive.

Furthermore, **factorMerger** offers yet another algorithm of hierarchical clustering. This is also an iterative procedure, but in each step it chooses model with the highest likelihood. While this algorithm is more complex than *DMR4glm* it is easily expandable for non-parametric models (using permutation tests instead of LRTs). The greedy algorithm is also available in two versions - comprehensive and sequential.

More detailed description of all algorithms implemented in **factorMerger** is given in the section *Algorithms overview*.

Algorithms overview

Sequential version

In the sequential version of the algorithm at the begining categorical variable is releveled. Depending on the model family chosen, we specify different statistics to set levels order.

model	metric
single dimensional gaussian	mean
multi dimensional gaussian	mean of isoMDS fit
binomial	success proportion
survival	relative survival coefficient

For single dimensional gaussian and binomial models groups are sorted by means and proportions of success, respectively. In survival case we estimate survival model which takes all factor levels separately. Then beta coefficient approximations specify levels order (base level gets coefficient equal to zero). Multi dimensional gaussian model needs additional preprocessing. We propose to order levels by means of isoMDS projection (into one dimension, currently isoMDS from package **MASS** is used). However, the projection is used only in this preliminary stage. In the merging phase of the algorithm all test statistics are calculated for multi dimensional gaussian model. Calculating isoMDS projection is an expensive procedure — it usually takes more time than the merging phase.

Having set the factor order, we may limit number of comparisons in each step.

Bibliography

Bondell, Howard D. and Brian J. Reich. 2008. "Simultaneous factor selection and collapsing levels in ANOVA." *Department of Statistics, North Carolina State University* .

Prochenka, Agnieszka. 2016*a*. *Delete or Merge Regressors algorithm*. chapter 4.3, p. 37.

Prochenka, Agnieszka. 2016*b*. *Delete or Merge Regressors algorithm*. chapter 5–6, pp. 44–91.

Tibshirani, Robert, Michael Saunders, Saharon Rosset, Ji Zhu and Keith Knight. 2005. "Sparsity and smoothness via the fused lasso." *Journal of the Royal Statistical Society* pp. 01–108.

Tukey, John. 1949. "Comparing Individual Means in the Analysis of Variance." *BIOMETRICS* pp. 99–114.