factorMerger: A Set of Tools to Support Results From Post Hoc Testing

by Agnieszka Sitko and Przemysław Biecek

Abstract *ANOVA*-like statistical tests for differences among groups are available for almost a hundred years. However, for large number of groups the results from commonly used *post-hoc* tests are often hard to interpret. To deal with this problem, the factorMerger package constructs and plots the hierarchical relation for the considered factor. Such a hierarchical structure is derived based on the *Likelihood Ratio Test* and is presented with the *Factor Merger Tree* created with the **ggplot2** package (Wickham, 2009). The current implementation handles one-dimensional and multi-dimensional Gaussian models as well as binomial and survival models. In this article the methodological background is outlined and the main functionalities of the packege are illustrated using real-data examples.

Introduction

In this article we present the factorMerger package that enriches results from *ANOVA* tests. The *ANOVA* method verifies the null hypothesis that a variable of interest y has the same distribution in all groups that are being compared. If this null hypothesis is rejected a more detailed analysis of differences among categorical variable levels might be needed. The traditional approach is to perform *pairwise post hoc tests* in order to verify which groups differ significantly.

One may find implementations of the traditional *post hoc tests* in many *R* packages. Package **agricolae** (de Mendiburu, 2016) offers a wide range of them. It gives one of the most popular *post hoc test*, Tukey HSD test (`HSD.test`), its less conservative version — Student-Newman-Keuls test (`SNK.test`) or Scheffe test (`scheffe.test`) which is robust to factor imbalance. These parametric tests are based on Student's t-distribution, thus, are reduced to Gaussian models only. In contrasts, **multcomp** package (Hothorn et al., 2008) can be used with generalized linear models (function `glht`) as it uses general linear hypothesis. Similarly to the **multcomp**, some implementations that accept `glm` objects are also given in **car** (`linearHypothesis`, Fox and Weisberg, 2011) and **lsmeans** (Lenth, 2016).

However, an undeniable disadvantage of single-step *post hoc tests* is the inconsistency of their results. For a fixed significance level, it is possible that mean in group A does not differ significantly from the one in group B, similarly with groups B and C. At the same time the 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 the literature. First, J. Tukey proposed an iterative procedure of merging factor levels based on the 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 in testing. 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, p. 37) is also adjusted to generalized linear models. It directly uses the hierarchical clustering to gain hierarchical structure of a factor. At the beginning, *DMR4glm* calculates the likelihood ratio test statistics for models arising from pairwise merging of factor levels or deleting factor levels against the full model (the one with all groups included). Then it performs agglomerative clustering taking LRT statistic as a distance — each step of clustering is associated with a model with different factor structure. Experimental studies (Prochenka, 2016, p. 44–91) showed that the *Delete or Merge Regressors*'s performance is better than *CAS-ANOVA*'s when it comes to the model accuracy. *Delete or Merge Regressors*'s implementation may be found in the **DMR** package (Maj et al., 2013).

In this article we present a more direct approach to the problem of merging groups that are being compared. The factorMerger package offers an algorithm of hierarchical clustering of factors based on a backward iterative procedure. In each step it chooses a model with the highest *Likelihood Ratio Test* test p-value or, in other words, the highest likelihood. While this algorithm is more complex than *DMR4glm*, it maximizes the likelihood on the merging path[1]. What is more, it is easily expandable for non-parametric models (using permutation tests instead of *LRT*s).

In addition to the comprehensive algorithm which tries to merge all possible pairs of levels in

[1] Although it may be shown that the *DMR* algorithm is a consistent model selection method, its performance on smaller datasets is undefined. TODO....

a step, also a *successive version* is provided. In the *successive version* only levels which are relatively close can be merged (levels distance is dependent on the model chosen). While the basic approach (all vs. all comparisons) may result in a slightly better partition from the statistical point of view, proposed extension (all vs. subsequent comparisons) seems to be more graceful when it comes to the interpretation. Moreover, the former algorithm is more computationally expensive.

Furthermore, the factorMerger package gives an approximate implementation of *DMR4glm* (skipping the deleting procedure) with its *successive version*.

More detailed description of all algorithms implemented in factorMerger is given in the section Methodology.

Methodology

Merging procedures implemented in the factorMerger package begin with the full model — with all levels of a given factor included — and iteratively merge one pair of levels until the factor is constant. Uniting two groups reduces by one the number of subsets, so, as initially we have finite number of levels, the procedure will eventually obtain one-level-factor and terminate. In a single iteration *all possible* pairs are considered and the one which optimizes some objective function is joined. Objective functions use likelihood-based statistics we will describe later on.

The factorMerger package gives the ability to perform analysis for the wide family of models and choose from the broad spectrum of merging approaches. Depending on the problem statement, some parts of the merging procedure may differ. The general sketch of the algorithm is described below.

Algorithm 1 The outline of the merging procedure

function MERGEFACTORS($response, factor, family, successive, method$)
  $pairsSet := generatePairs(response, factor, successive)$
  $\mathcal{M} := createModel(response, factor, family)$
  while $|levels(factor)| > 1$ **do**
    $toBeMerged := \text{argmax}_{\text{pair} \in \text{pairsSet}} objectiveFunction(pair)$
    $\mathcal{M} := updateModel(M_0,\ toBeMerged)$
    $factor := mergeLevels(factor, pair)$
    $pairsSet := removePair(pairsSet, pair)$
  end while
end function

Model family

In the current version the package supports parametric models:

- single dimensional Gaussian (with the argument `family = "gaussian"`),
- multi dimensional Gaussian — Gaussian model with multiple y outputs (with the argument `family = "gaussian"`)[2],
- binomial (with the argument `family = "binomial"`),
- survival (with the argument `family = "survival"`).

Each case has its own method of estimating model parameters and a specific likelihood formula.

[2]Both single dimensional and multi dimensional Gaussian models use `family = "gaussian"`. However, multi dimensional model uses different functions for likelihood estimation and may require additional preprocessing, thus, it is considered as a separate category.

Single dimensional Gaussian model Here we consider the following model.

$$y = X\beta + \epsilon, \;\; \epsilon \sim \mathcal{N}\left(0, \sigma^2\right),$$

where X is a binary matrix responsible for encoding group membership.

Under the above assumption, denoting sample size as n, we may formulate the likelihood of the Gaussian linear model (Friedman et al., 2001, p.31)

$$L\left(\beta, \sigma | y\right) = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left(-\frac{1}{2}\left(y - X\beta\right)^T \left(y - X\beta\right)/\sigma^2\right)$$

and its logarithm

$$l\left(\beta, \sigma | y\right) = -\frac{n}{2}\log\left(2\pi\right) - \frac{n}{2}\log\left(\sigma^2\right) - \frac{1}{2}\left(y - X\beta\right)^T \left(y - X\beta\right)/\sigma^2.$$

To calculate the loglikelihood we use `logLik.lm{stats}`.

Multi dimensional Gaussian model Here we consider the model.

$$Y = X\beta + E, \;\; E \sim \mathcal{N}(0, \Sigma),$$

where X is a binary matrix responsible for encoding group membership, $Y = (y_1, y_2, ..., y_k)$ is a k-dimensional response and $E = (\epsilon_1, \epsilon_2, ..., \epsilon_k)$ is a k-dimensional error.

Having the sample size denoted as n, we may calculate the likelihood

$$L\left(\beta, \Sigma | Y\right) = \left(|2\pi\Sigma|\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\left(Y - X\beta\right)^T \Sigma^{-1} \left(Y - X\beta\right)\right)$$

and its logarithm

$$l\left(\beta, \sigma | Y\right) = -\frac{n}{2}\log\left(2\pi\right) - \frac{n}{2}\log\left(|\Sigma|\right) - \frac{1}{2}\left(Y - X\beta\right)^T \Sigma^{-1} \left(Y - X\beta\right).$$

Unfortunately, **stats** or any commonly used *R* package do not support multiple responses in the loglikelihood calculation for linear Gaussian models. In the package we use `logLik.lm` implementation introduced in the **Atools** package (B<b6>ck, 2014) and the `dmvnorm{mvtnorm}` (Genz and Bretz, 2009) implementation for multivariate normal density estimation.

Binomial model In the binomial case we assume that

$$y \sim \mathcal{B}\left(p, n\right)$$

where $\mathcal{B}\left(p, n\right)$ is the binomial distribution with probability of success p and number of trials n. We consider the logit model

$$\ln\left(\frac{p}{1-p}\right) = X\beta$$

with X – binarized matrix reprezentation of a factor.

Let $z = \sum_{i=1}^{n} y_i$. We may write the likelihood as follow (Czepiel, 2002)

$$L\left(\beta | y\right) = \frac{n!}{z!\left(n-z\right)!} p^z \left(1-p\right)^{n-z}.$$

Thus, the logarithm of the likelihood may be expressed as follow

$$l\left(\beta | y\right) = zX\beta - n\log\left(1 + \exp^{X\beta}\right).$$

TODO: Policzyć loglik, czy na pewno dobrze.

To calculate loglikelihood for the binomial model we use `logLik.glm{stats}`.

Survival model

Pairs considered

Set of hypotheses that are tested during merging may be either comprehensive or limited. This gives two possibilities:

- *all-to-all* (with the argument `successive = FALSE`),
- *successive* (with the argument `successive = TRUE`).

The version *all-to-all* considers all possible pairs of factor levels. In the *successive* approach factor levels are preliminarily sorted and then only consecutive groups are tested for means equality.

Defining levels similarity

The factorMerger package for each model family and merging strategy implements two types of a single iteration of the algorithm. They use one of the following:

- *Likelihood Ratio Test* (with the argument `method = "LRT"`),
- *agglomerative clustering with constant distance matrix* (based on the *DMR4glm* algorithm, with the argument `method = "hclust"`).

The *successive* merging

In the *successive* version of the algorithm levels of a categorical variable are sorted. The order depends on the model chosen family chosen.

model	metric
one-dimensional Gaussian	average
multi-dimensional Gaussian	average of the isoMDS transformation
binomial	proportion of successes
survival	relative survival rate

Table 1: Factor ordering by model family

For one-dimensional Gaussian and binomial models groups are sorted by means and proportions of success, respectively. In the 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. First, group means are computed. Then they are projected into one dimension with the use of the Kruskal's non-metric multidimensional scaling. The factorMerger uses `isoMDS` implementation from the package **MASS** (Venables and Ripley, 2002).

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

The *all-to-all* merging

Short description...

The Likelihood Ratio Test statistics

The substantial part of factorMerger's algorithms is calculating the *Likelihood Ratio Test* statistics. In this section we define *LRT* statistic used in merging.

Let us assume y is a response variable and C is a factor with k levels ($C \in \{1, 2, ..., k\}$). We denote as h some linear hypothesis on the levels of C, M_0 the initial model (taking all factor levels independently) and M_h — the model under h. Then, $LRT(M_h|M_0)$ statistic based on the *Likelihood Ratio Test* is defined as below.

$$LRT(M_h|M_0) = 2 \cdot l(M_0) - 2 \cdot l(M_h),$$

where $l(\cdot)$ is log-likelihood function.

As M_h is nested in M_0, the likelihood of M_h is not greater than the M_0's likelihood. Therefore, if $\mathcal{H}$ is a set of considered linear hypothesis, hypothesis

$$\text{argmin}_{h\in\mathcal{H}} LRT(M_h|M_0) = \text{argmax}_{h\in\mathcal{H}} l(M_h)$$

will reduce likelihood the least.

Asymptotic behaviour of the *LRT* statistic A convenient result by Samuel S. Wilks (Wilks, 1938) shows that $LRT(M_h|M_0)$ tends asymptotically to chi-squared distribution with degrees of freedom equal to the difference in degrees of freedom between M_0 and M_h as number of observations approaches infinity. This convergence will be used to evaluate model's 'statistical correctness'.

The *Likelihood Ratio Test*-based merging

The *Likelihood Ratio Test*-based approach minimizes likelihood reduction in the merging path. It may be summarized as follow.

TODO: Rozwinąć... (Analogia do LRT testów, ale można uprościć do samego loglik)

Algorithm 2 Merging with the *LRT*

```
function MERGEFACTORS(response, factor, successive)
    pairsSet := generatePairs(response, factor, successive)
    M_0 := full model
    while levels(factor) > 1 do
        toBeMerged := argmax_{pair∈pairsSet} l(updateModel(M_0, pair))
        M_0 := updateModel(M_0, toBeMerged)
        factor := mergeLevels(factor, pair)
        pairsSet := pairsSet \ pair
    end while
end function
```

The *DMR4glm*-based merging

TODO: Wstępny opis

Algorithm 3 Merging with agglomerative clustering

```
function MERGEFACTORS(response, factor, successive)
    pairsSet := generatePairs(response, factor, successive)
    dist := set of distances
    for all pair ∈ pairsSet do
        h := {μ_{pair_1} = μ_{pair_2}}        ▷ hypothesis under which pair is merged
        dist[pair] = LRT(M_h|M_0)
    end for
    if successive then
        hClust(dist, method = "single")
    else
        hClust(dist, method = "complete")
    end if
end function
```

An *R* package factorMerger

The factorMerger package provides easy-to-use functions for factor merging and visualizing obtained results.

TODO: Tutaj da się opis głównych funkcji - co robią i tabelki z opisem parametrów

Merging and getting results

Visualizations

TODO: Tutaj dojdzie jeszcze opis każdego z paneli osobno

CASE STUDY: PISA2012

Summary

Some summary.

Acknowledgements

The authors thank to ... .

Work on this package was financially supported by the *NCN Opus grant 2016/21/B/ST6/02176*.

Bibliography

A. B<b6>ck. *Atools: Atools*, 2014. URL https://R-Forge.R-project.org/projects/biostat/. R package version 0.2/r191. [p3]

H. D. Bondell and B. J. Reich. Simultaneous factor selection and collapsing levels in ANOVA. *Department of Statistics, North Carolina State University*, 2008. [p1]

S. A. Czepiel. Maximum likelihood estimation of logistic regression models: theory and implementation. *Available at czep. net/stat/mlelr. pdf*, 2002. [p3]

F. de Mendiburu. *agricolae: Statistical Procedures for Agricultural Research*, 2016. URL https://CRAN.R-project.org/package=agricolae. R package version 1.2-4. [p1]

J. Fox and S. Weisberg. *An R Companion to Applied Regression*. Sage, Thousand Oaks CA, second edition, 2011. URL http://socserv.socsci.mcmaster.ca/jfox/Books/Companion. [p1]

J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2001. [p3]

A. Genz and F. Bretz. *Computation of Multivariate Normal and t Probabilities*. Lecture Notes in Statistics. Springer-Verlag, Heidelberg, 2009. ISBN 978-3-642-01688-2. [p3]

T. Hothorn, F. Bretz, and P. Westfall. Simultaneous inference in general parametric models. *Biometrical Journal*, 50(3):346–363, 2008. [p1]

R. V. Lenth. Least-squares means: The R package lsmeans. *Journal of Statistical Software*, 69(1):1–33, 2016. doi: 10.18637/jss.v069.i01. [p1]

A. Maj, A. Prochenka, and P. Pokarowski. *DMR: Delete or Merge Regressors for linear model selection.*, 2013. URL https://CRAN.R-project.org/package=DMR. R package version 2.0. [p1]

A. Prochenka. *Delete or Merge Regressors algorithm*, chapter 4.3, 5–6, pages 37, 44–91. 2016. [p1]

R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society*, pages 01–108, 2005. [p1]

J. Tukey. Comparing Individual Means in the Analysis of Variance. *BIOMETRICS*, pages 99–114, 1949. [p1]

W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. URL http://www.stats.ox.ac.uk/pub/MASS4. ISBN 0-387-95457-0. [p4]

H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009. ISBN 978-0-387-98140-6. URL http://ggplot2.org. [p1]

S. S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62, 1938. [p5]

Agnieszka Sitko
University of Warsaw
Faculty of Mathematics, Informatics and Mechanics
Poland
ag.agnieszka.sitko@gmail.com

Przemysław Biecek
University of Warsaw
Institute of Applied Mathematics and Mechanics
Poland
P.biecek@mimuw.edu.pl