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

**Agnieszka Sitko** *Faculty of Mathematics, Informatics and Mechanics, University of Warsaw*
**Przemysław Biecek** *Faculty of Mathematics, Informatics and Mechanics, University of Warsaw*

*ANOVA*-like statistical tests for differences among groups are available for almost a hundred years. But 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 among compared groups. Such hierarchical structure is derived based on the *Likelihood Ratio Test* and is presented with the *Merging Paths Plots* created with the **ggplot2** package. The current implementation handles one-dimensional and multi-dimensional Gaussian models as well as binomial and survival models. This article presents the theory and examples for a single-factor use cases.
*Package webpage*: https://github.com/geneticsMiNIng/FactorMerger

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

## Introduction

In this article we present a **factorMerger** package that enriches results from *ANOVA* tests. The *ANOVA* method verifies the null hypothesis that the 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 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, Bretz and Westfall, 2008) can be used with generalized linear models (function `glht`) as it uses general linear hypothesis. Similarly to **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 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 literature. 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 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 initial 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 *Delete or Merge Regressors*'s performance is better than *CAS-ANOVA*'s when it comes to model accuracy.

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 base on an iterative procedure. In each step it chooses model with the highest p-value from *LRT* test. While this algorithm is more complex than *DMR4glm*, it maximizes likelihood on the merging path. What is more, it is easily expandable for non-parametric models (using permutation tests instead of LRTs). This algorithm is also available in two versions - comprehensive and sequential.

Furthermore, **factorMerger** package gives also an approximate implementation of *DMR4glm* (skipping the deleting procedure). In addition to the base algorithm, it also provides its sequential version. It 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 extension (all vs. subsequent comparisons) seems to be more graceful when it comes to the interpretation. Moreover, the former is more computationally expensive.

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

### Algorithms overview

The **factorMerger** package gives a user the ability to perform analysis for the wide family of models and choose from the broad spectrum of merging approaches.

In the current version the package supports parametric models:

- one-dimensional Gaussian (with the argument `family="gaussian"`),
- multi dimensional Gaussian (with the argument `family="gaussian"`),
- binomial (with the argument `family="binomial"`),
- survival (with the argument `family="survival"`).

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

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

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

The **factorMerger** package also implements two strategies of a single iteration of the algorithm. They use one of the following:

- *Likelihood Ratio Test*,
- *agglomerative clustering with constant distance matrix* (based on the *DMR4glm* algorithm).

*Sequential version*

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

Table 1: Factor ordering by model family

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

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

*Likelihood Ratio Test*

The substantial part of **factorMerger** 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, the *Likelihood Ratio Test* statistic is calculated as a logarithm of $M_0$ and $M_h$ likelihood ratio

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

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.

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'.

*Agglomerative clustering*

The `hclust`-based approach is an approximation of *DMR4glm*. The algorithm process is described below.

**Algorithm 1** Merging with agglomerative clustering

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

*Greedy algorithm*

In contrary to the previous method, *greedy* approach minimizes likelihood reduction in each step. It may be summarized as follow.

**Algorithm 2** Merging with *LRT*

**function** MERGEFACTORS($response, factor, subsequent$)
  $pairsSet := generatePairs(response, factor, subsequent)$
  $M_0 :=$ full model
  **while** $levels(factor) > 1$ **do**
    $toBeMerged = \text{argmax}_{pair \in pairsSet} l(updateModel(M_0, pair))$
    $M_0 := updateModel(M_0, toBeMerged)$
    $factor := mergeLevels(factor, pair)$
    $pairsSet := pairsSet \setminus pair$
  **end while**
**end function**

### The *R* package factorMerger

The **factorMerger** package provides easy-to-use functions for factor merging and visualizing obtained results. Package's functionalities are illustrated using 3-dimensional Gaussian response and factor variable with 5 levels.

```
library(factorMerger)
sample <- generateMultivariateSample(100, 5, 3)
fm <- mergeFactors(response = sample$response, factor = sample$factor,
```

```
                    family = "gaussian", subsequent = TRUE,
                    method = "LRT", penalty = 2)
```

`mergeFactors` takes arguments: `response` – vector/matrix of response (note: in survival model response must be of a class `Surv`), `factor` – factor to be merged, `family` – model family, `subsequent` – binary variable specifying which levels are permitted to be merged, `method` – algorithm step method (either "LTR" or "hclust"), `penalty` – penalty used in GIC calculations.

Sample results of `mergeFactors` in the multinomial Gaussian example are presented in the Table 2.

Table 2: Multinomial Gaussian merging results

| groupA | groupB | loglikelihood | p-value | GIC |
|---|---|---|---|---|
| | | -519.4360 | 1.0000 | 1048.872 |
| (A) | (B) | -522.1138 | 0.1713 | 1052.227 |
| (E) | (D) | -524.8228 | 0.1184 | 1055.646 |
| (C) | (E)(D) | -526.9399 | 0.1202 | 1057.880 |
| (C)(E)(D) | (A)(B) | -534.2032 | 0.0056 | 1070.406 |

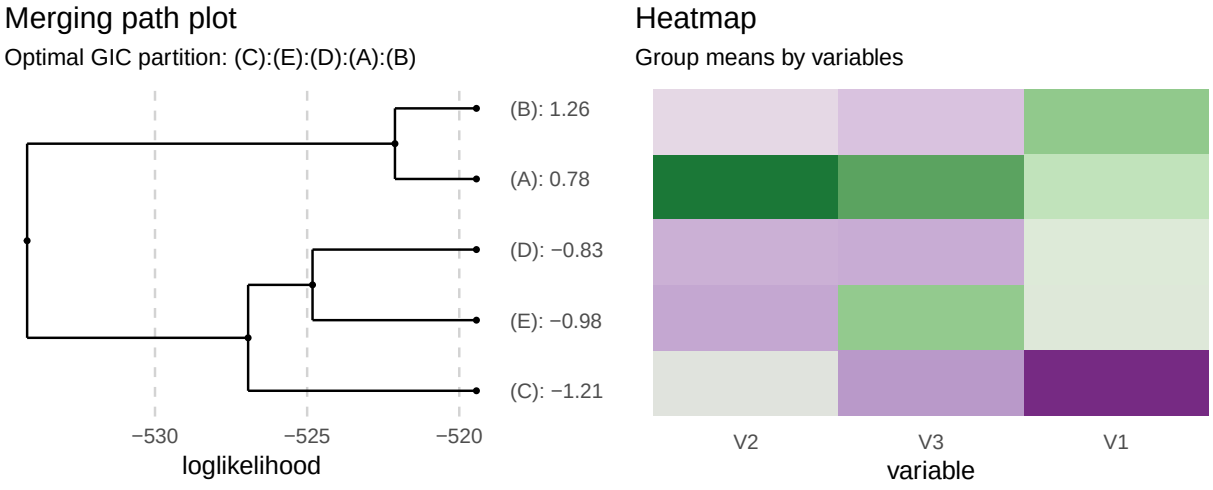


Figure 1: An example for a multi-dimensonal Gaussian reposonse with corresponding heatmap

The **factorMerger** package gives plenty of possibilities to plot merging results. We may want to plot cluster tree in a simplified form (nodes are distributed evenly) or customized (nodes represent group statistic). We can choose between plotting p-value on the x axis or loglikelihood. We can also decide if we want to mark the best model in GIC criterion. There are also many possibilities of summarizing response variable visually.

In Figures 1 and 2 each interval in the OX axis corresponds to the 0.95 quantile of chi-square distribution with one degree of freedom. Models distant more than this interval may be considered as significantly different.

Find more examples and visualizations in the **factorMerger** vignette.

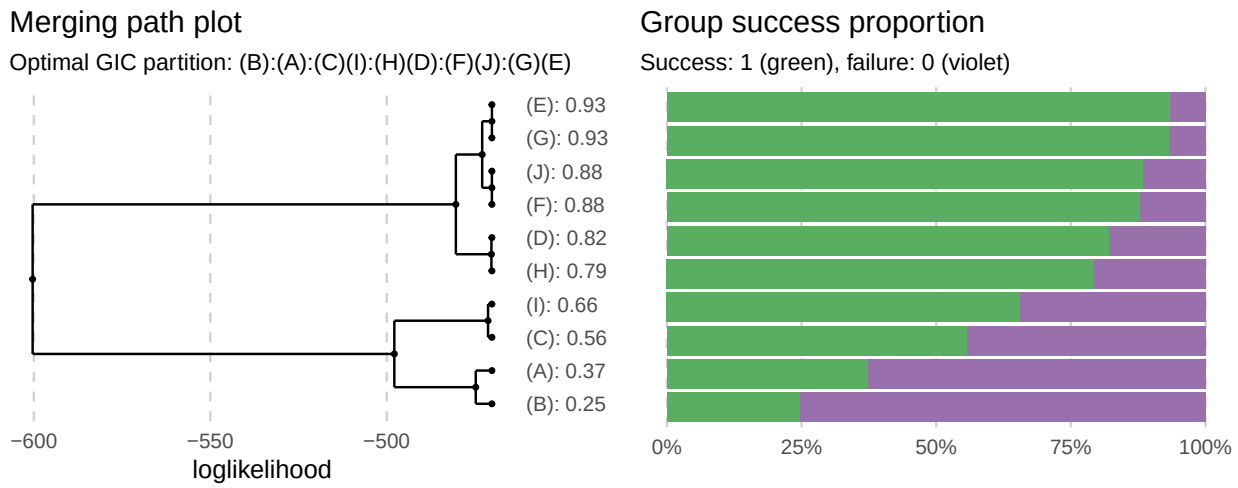


Figure 2: An example for a binomial reposonse

**Bibliography**

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

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

Fox, John and Sanford Weisberg. 2011. *An R Companion to Applied Regression*. Second ed. Thousand Oaks CA: Sage.
**URL:** *http://socserv.socsci.mcmaster.ca/jfox/Books/Companion*

Hothorn, Torsten, Frank Bretz and Peter Westfall. 2008. "Simultaneous Inference in General Parametric Models." *Biometrical Journal* 50(3):346–363.

Lenth, Russell V. 2016. "Least-Squares Means: The R Package lsmeans." *Journal of Statistical Software* 69(1):1–33.

Prochenka, Agnieszka. 2016. *Delete or Merge Regressors algorithm*. chapter 4.3, 5–6, pp. 37, 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.

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