University of Warsaw
Faculty of Mathematics, Informatics and Mechanics

Agnieszka Sitko
Student no. 339398

Merge and Select: Visualization of a likelihood based k-sample adaptive fusing and model selection

Master's thesis
in MATHEMATICS

Supervisor:
dr hab. Przemysław Biecek
Faculty of Mathematics, Informatics and Mechanics
University of Warsaw

September 2017

Supervisor's statement

Hereby I confirm that the presented thesis was prepared under my supervision and that it fulfils the requirements for the degree of Master of Mathematics.

Date Supervisor's signature

Author's statement

Hereby I declare that the presented thesis was prepared by me and my supervisor, Przemysław Biecek. None of its contents was obtained by means that are against the law.

The thesis has never before been a subject of any procedure of obtaining an academic degree.

Moreover, I declare that the present version of the thesis is identical to the attached electronic version.

Date Author's signature

Abstract

In this thesis we introduce Merge and Select — a methodology — and **factorMerger** — an R package — for exploration and visualization of k-group comparisons. Comparison of k-groups is one of the most important issues in exploratory analyses and it has zillions of applications. The classical solution is to test a null hypothesis that observations from all groups come from the same distribution. If the global null hypothesis is rejected a more detailed analysis of differences among pairs of groups is performed. The traditional approach is to use pairwise *post hoc tests* in order to verify which groups differ significantly. However, this approach fails with large number of groups in both interpretation and visualization layer. The Merge and Select methodology solves this problem by using easy to understand description of LRT based similarity among groups.

Keywords

post-hoc testing, hierarchical clustering, likelihood ratio test

Thesis domain (Socrates-Erasmus subject area codes)

11.2 Statistics

Subject classification

62 Statistics
62J Linear inference, regression
62J15 Paired and multiple comparisons

Tytuł pracy w języku polskim

Merge and Select: Wizualizacja adaptacyjnego łączenia populacji w oparciu o wiarygodność modelu

Contents

Chapter 1

Introduction and Motivation

One of the most frequent tasks in exploratory analyses is the comparison of k groups. There are zillions of applications, like comparisons of different medical treatments, comparisons of different countries or comparisons of segments of clients. The classical solution is to test the global hypothesis, that all groups are equal. If the global null hypothesis is rejected a more detailed analysis of differences among pairs of groups is needed. The traditional approach is to perform *post hoc tests* in order to verify which groups differ significantly.

As we show later, this approach fails if the number of groups is large as the number of pairs quickly grows beyond easy interpretation.

The larger the number of groups, the more pronounced is the problem with classical post-hoc testing. For example, in the PISA study (*Program for International Students Assessment* 2012) data about academic performance of 15 year old kids from 65 countries is collected. One can use tests like ANOVA or other k-sample tests to verify whatever there are any differences between countries but then the question arises how counties are different. The total number of pairwise comparison is $\frac{65(65-1)}{2} = 2080$ and obviously it is not easy to present such a number of results in an easy to understand way. Figure 1.1, where results for only 11 European countries are presented, shows how hard it is to read anything when the number of groups is not small.

The problem with post-hoc testing is also related to inconsistency of results. For a fixed significance level, it is possible that the 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.

To deal with these problems, we introduce the *Merge and Select* methodology along with a tool — **factorMerger** — a library for R software (R Core Team 2017). The aim of the methodology is to enrich results from *k-sample* tests together with providing the variety of plots designed for deeper understanding analyzed models. An example of a Merge and Select's visualization is presented in Figure 1.2.

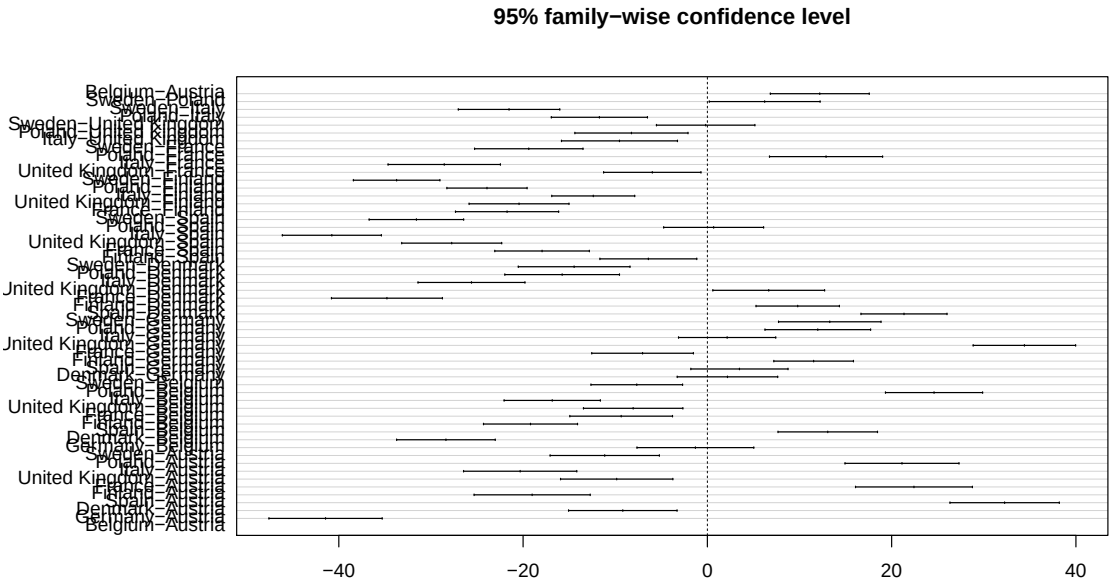


Figure 1.1: Classical approach to graphical presentation of post-hoc testing of 11 groups with the use of the `plot.tukeyHSD` function. For each pair of countries the plot presents average difference and a 95% confidence interval. Based on data from PISA 2012 study for 11 counties (55 pairs).

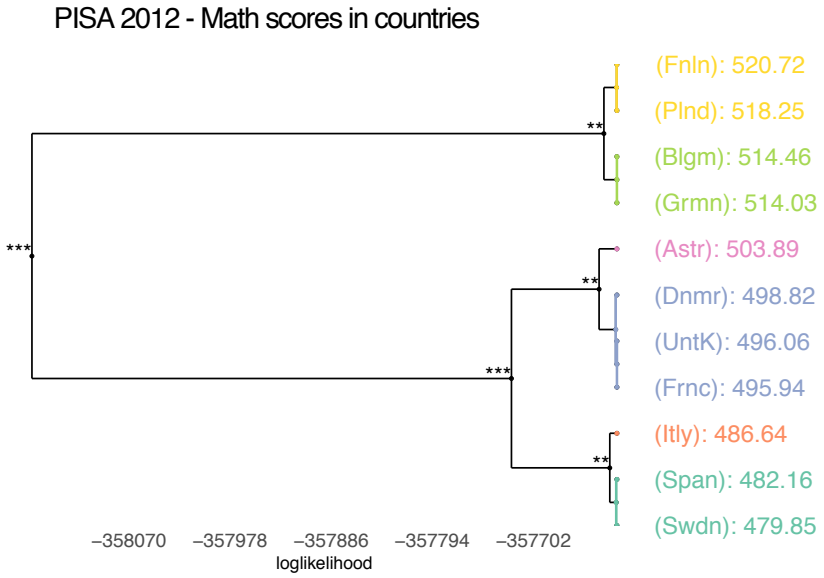


Figure 1.2: Merging Paths Plot for academic performance among 11 countries. The same data as in the Figure 1.1 is used. Such a plot is much easier to read and interpret.

The aim of **factorMerger** package is to provide informative and easy to understand visualizations of post-hoc comparisons. It works for wide spectrum of probability distribution families of dependent variable as it is based on linear model with Gaussian outcome, generalized linear model with binomial outcome and Cox proportional hazard model with censored outcome. Merge and Select's visualizations, *Merging Path Plots* (MPP, an example of MPP is presented in Figure 1.2), show consistent and non-overlapping adaptive fusing of groups based on the likelihood ratio test (LRT) statistics. In addition, the Generalized Information Criterion (GIC) is presented for fused models. This criterion may be used to choose the optimal fusion of groups.

Chapter 2

Background and Related Work

One may find implementations of the traditional *post hoc tests* in many *R* packages (R Core Team 2017). For example, package **agricolae** (de Mendiburu 2016) offers a wide range of them. It gives one of the most popular *post hoc test*, Tukey HSD test (function `HSD.test`), its less conservative version — Student-Newman-Keuls test (function `SNK.test`) or Scheffe test (function `scheffe.test`) which is robust to group imbalance. These parametric tests are based on Student's t-distribution, thus, are reduced to Gaussian models only. In contrast, **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 & Weisberg 2011) and **lsmeans** (Lenth 2016).

But what about the problem of clustering categorical variable into non-overlapping groups? It 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, again, statistical test used in this approach made it limited to Gaussian models.

Collapse And Shrinkage in ANOVA (*CAS-ANOVA*, Bondell & 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 constraints taken on pairwise differences within a factor, which yields to their smoothing. Yet another approach that is also adjusted to generalized linear models is presented by *Delete or Merge Regressors* algorithm (*DMR4glm*, Maj-Kańska et al. 2015). It directly uses the agglomerative hierarchical clustering (Peter Rousseeuw & Leonard Kaufman 1990) to build a hierarchical structure of groups that are being compared. Experimental studies (Maj-Kańska et al. 2015) show that *Delete or Merge Regressors*'s performance is better than *CAS-ANOVA*'s when it comes to the accuracy of the resulting model. The *Delete or Merge Regressors* method was first implemented in the **DMR** R package (Maj et al. 2013) and is reimplemented for broader number of model families in the **factorMerger** package.

The approach presented in this thesis extends approaches presented above in following ways:

- in comparison to pairwise tests Merge and Select results are easier to interpret.
- the **factorMerger** visualizations are created based on **ggplot2** (Wickham 2009) graphics and is easy to customize.
- in comparison to Fused LASSO, Merge and Select is based on the likelihood ratio test statistic, which has known asymptotic properties. This allows to calculate p-values for selected pairs of groups.

- in comparison to Fused LASSO obtained group effects of Merge and Select (like averages) are not biased and are easier to interpret.
- in comparison to **DMR** the modeling can be applied to wider variety of regression models, like generalized linear models and survival regression models.
- as we will show later, in comparison to **DMR** the resulting structure of groups in Merge and Select is more stable.

In the next chapter we will present the methodology beyond the **factorMerger** package.

Chapter 3

Merging Path Plots

Let k stands for the number of groups, while n_i stands for the number of observations in group $i \in \{1, ..., k\}$. Let y_{ij} denote an observed value of variable of interest for observation $j \in \{1, ..., n_i\}$ in group i. We assume that $y_{ij} \sim F(\theta_i)$, where F is a distribution from exponential family parametrized by $\theta \in \Theta$.

The global null hypothesis is

$$H_0 : \forall_{i \in \{1,...,k\}} \theta_i = \theta_1$$

and can be tested with the Likelihood Ratio Test for k samples. If the global null hypothesis is rejected, then in the post-hoc analysis we are looking for groups with equal distributions, that is sets of indexes J such as $\forall_{i,j \in J} \theta_i = \theta_j$

In Merge and Select these sets are obtained in an iterative fashion. In every step two groups are merged into a single one. This step is repeated as long as there is more than one group. The general sketch of the algorithm is described below.

The merging procedure begins with a full model — with all original groups — and iteratively merges a pair of groups until all of them are combined. For considered families of distributions we use generalized linear models or Cox proportional hazard model. Each merging of two groups reduces by one the number of degrees of freedom of a model. In a single iteration pairs *worth fusing* are considered and the one which optimizes an objective function is merged. In general any model statistic may be used as an objective function, but here we are using the likelihood statistic. We will specify it in details in the next chapter. A general formulation of the merging procedure is described in Algorithm 1.

Algorithm 1 The outline of Merge and Select algorithm implemented in **factorMerger**

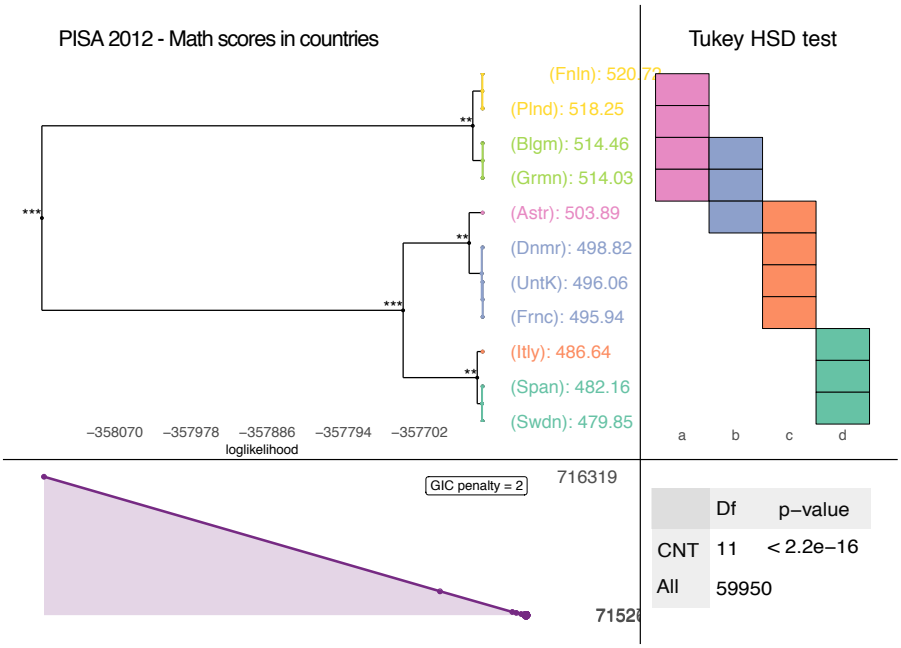
function MERGEFACTORS($responseVariable, groupingVariable, adjacent$)
  $currentModel := createModel(responseVariable, groupingVariable)$
  $mergingPath := list(currentModel)$
  while $|levels(groupingVariable)| \geq 1$ **do**
    $pairsSet := generatePairs(groupingVariable, responseVariable, adjacent)$
    $selectedPair := \text{argmax}_{\text{pair} \in \text{pairsSet}} objectiveFunction(pair, responseVariable, groupingVariable)$
    $groupingVariable := mergeLevels(groupingVariable, selectedPair)$
    $currentModel := createModel(responseVariable, groupingVariable)$
    $mergingPath := add(mergingPath, currentModel)$
  end while
  return(mergingPath)
end function

The result of the Algorithm 1 is a list of k shrinking models $\mathcal{M}_i$, where $i \in \{1, ..., k\}$. In **factorMerger** these models are presented in a graphical way in a *Merging Path Plot* along with diagnostic criteria like Generalized Information Criteria and other graphical summaries. Merge and Select's plot contains four panels, that encapsulate all important information in a compact form. An example of these panels is presented in the Figure 3.1.

The statistics presented in this plot are described in following sections.

A. The Merging Paths Plot panel
Shows the hierarchical structure of similarity between groups.
Stars presents how significant are differences between two clusters.
Height of the join corresponds to the likelihood of the combined model.

B. The response panel
Shows graphical summaries for group separately.
Use the responsePanel argument to select the desired summary.



C. The GIC panel
Shows the *Generalized Information Criterion* for all models in MPP panel.
Default *penalty=2* corresponds to the AIC criterion.
Colours in the MPP panel corresponds to the optimal model.

D. The summary panel
Shows the no. groups, no. observations,
and the p-value for global hypothesis
that parameter of interest is equall in all groups.

Figure 3.1: Four panels of the **factorMerger**'s visualization for the PISA dataset for 11 countries. Panel A summarizes the structure of group similarities. It shows the list of models returned in the Algorithm 1. The OX axis presents values of log-likelihood function for each model from the list. Labels on the right margin present averages of variable of interest for different groups. Stars in different joins of the tree summarize pairwise tests for selected groups of variables. Panel B summarizes the distribution of variable of interest in each group. The summary plotted in this panel may be changed depending on the model family. Panel C shows the Generalized Information Criteria for each model from the list. Panel D presents results from the test for global null hypothesis. Colors in panels A and B are consistent and correspond to an optimal segmentation of groups based on the GIC score.

3.1. Model families

The Merge and Select algorithm can be performed for any likelihood based model. Current version of the **factorMerger** package supports following parametric models:

- one-dimensional Gaussian (with the argument `family = "gaussian"`). Here

$$y_{ij} \sim \mathcal{N}\left(\mu_j, \sigma^2\right)$$

  and corresponding logarithm of likelihood

$$l\left(\mu, \sigma | y\right) = -\frac{n}{2}\log\left(2\pi\right) - \frac{n}{2}\log\left(\sigma^2\right) - \sum_{j=1}^{k}\sum_{i=1}^{n_j}\frac{1}{2}\left(y_{ij} - \mu_j\right)^2/\sigma^2.$$

  Group summaries are averages – maximum likelihood estimates for μ_j.

- n-dimensional Gaussian (with the argument `family = "gaussian"`). Here Y_{ij} and M_j are vectors and

$$Y_{ij} \sim \mathcal{N}\left(M_j, \Sigma\right).$$

  The corresponding logarithm of likelihood function

$$l\left(M, \Sigma | Y\right) = -\frac{n}{2}\log\left(2\pi\right) - \frac{n}{2}\log\left(|\Sigma|\right) - \sum_{j=1}^{k}\sum_{i=1}^{n_j}\frac{1}{2}\left(Y_{ij} - M_j\right)^T\Sigma^{-1}\left(Y_{ij} - M_j\right).$$

  Note that, both one-dimensional and n-dimensional Gaussian models use `family = "gaussian"`. However, the visual summary of n-dimensional data requires additional preprocessing – dimensionality reduction – thus, it is considered as a separate category. Group summaries are averages.

- binomial (with the argument `family = "binomial"`). Here

$$y_{ij} \sim \mathcal{B}\left(p_j, 1\right).$$

  After adding the logit link function

$$\log\left(\frac{p_j}{1 - p_j}\right) = \beta_j$$

  one may write the logarithm of likelihood

$$l\left(\beta | y\right) = \sum_{j=1}^{k}\sum_{i=1}^{n_j} y_{ij}\beta_j - y_{ij}\log\left(1 + \exp\beta_j\right) + \left(1 - y_{ij}\right)\log\left(1 + \exp\beta_j\right).$$

  Group summaries are proportions of successes as estimates of p.

- survival (with the argument `family = "survival"`). Here we consider the *Cox proportional hazard model* (Cox 1992). Let $\lambda_0(t)$ be the baseline hazard function, where t denotes time. Then the hazard function for group j may be expressed as

$$\lambda_j(t) = \lambda_0(t) \cdot \exp(\alpha_j).$$

Corresponding logarithm of partial likelihood is

$$l\left(\alpha|\, y\right) = \sum_{i,j:C_{ij}=1} \left(\alpha_j - \log \left(\sum_{kl:y_{kl} \geq y_{ij}} \exp\left(\alpha_k\right) \right) \right).$$

where C_{ij} is the censoring status, $C_{ij} = 1$ means that the observation i from group j is not censored. For this model hazard ratios are the group summaries.

The fusing algorithm that is used in Merge and Select is based on the Likelihood Ratio Test statistic defined as

$$LRT(\mathcal{M}_1; \mathcal{M}_2) = 2 \cdot l(\widehat{\beta_{\mathcal{M}_2}}|y) - 2 \cdot l(\widehat{\beta_{\mathcal{M}_1}}|y), \tag{3.1}$$

where $\mathcal{M}_1$ and $\mathcal{M}_2$ are two nested models. Each model corresponds to a grouping of observations. Groupings for both models are equal except that two groups in $\mathcal{M}_2$ are merged in one group in $\mathcal{M}_1$. The higher the $LRT(\mathcal{M}_1; \mathcal{M}_2)$, the more different are the merged groups. One may interpret the $LRT(\mathcal{M}_1; \mathcal{M}_2)$ as a distance between groups for model $\mathcal{M}_1$ and $\mathcal{M}_2$.

The advantage of the *LRT statistic* is the known asymptotic behavior (see Wilks 1938). For nested models $\mathcal{M}_2$ a $\mathcal{M}_1$ that differs by one degree of freedom it holds

$$LRT(\mathcal{M}_1; \mathcal{M}_2) \overset{n \to \infty}{\sim} \chi^2_1.$$

This asymptotic distribution is used in **factorMerger** to present statistical significance of group joins with the argument `panelGrid = TRUE` of the `plot.factorMerger` function.

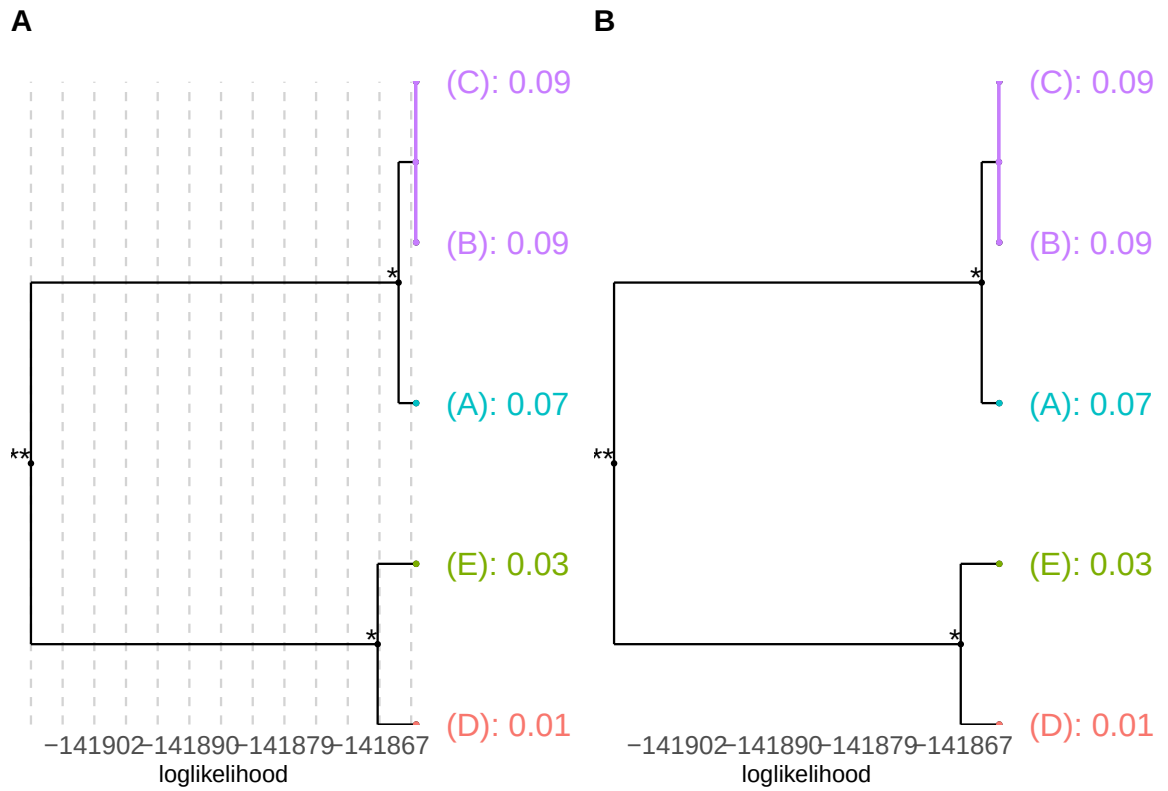


Figure 3.2: Merging Path Plots with panel grid (Panel A) and without panel grid (Panel B). In Panel A each interval in the OX axis corresponds to the 0.95 quantile of chi-square distribution with one degree of freedom. Models distant more than the length of this interval may be considered as significantly different

3.2. Group summaries

The right panel of the visualization shows graphical summaries of variable of interest in groups. Use the `responsePanel` argument to choose how groups shall be presented.

Available options are showed in the Figure 3.3. Depending on the family of the variable of interest different summaries are appropriate. Possible combinations are denoted in the Table 3.1.

responsePanel	family gaussian	binomial	survival
frequency	+	+	+
means	+		
boxplot	+		
tukey	+		
heatmap	+		
profile	+		
proportion		+	
survival			+

Table 3.1: Different types of graphical summary are appropriate for different model families. Pluses denote which `responsePanel` may be used for which model family. Examples for each type of panel are presented in the Figure 3.3.

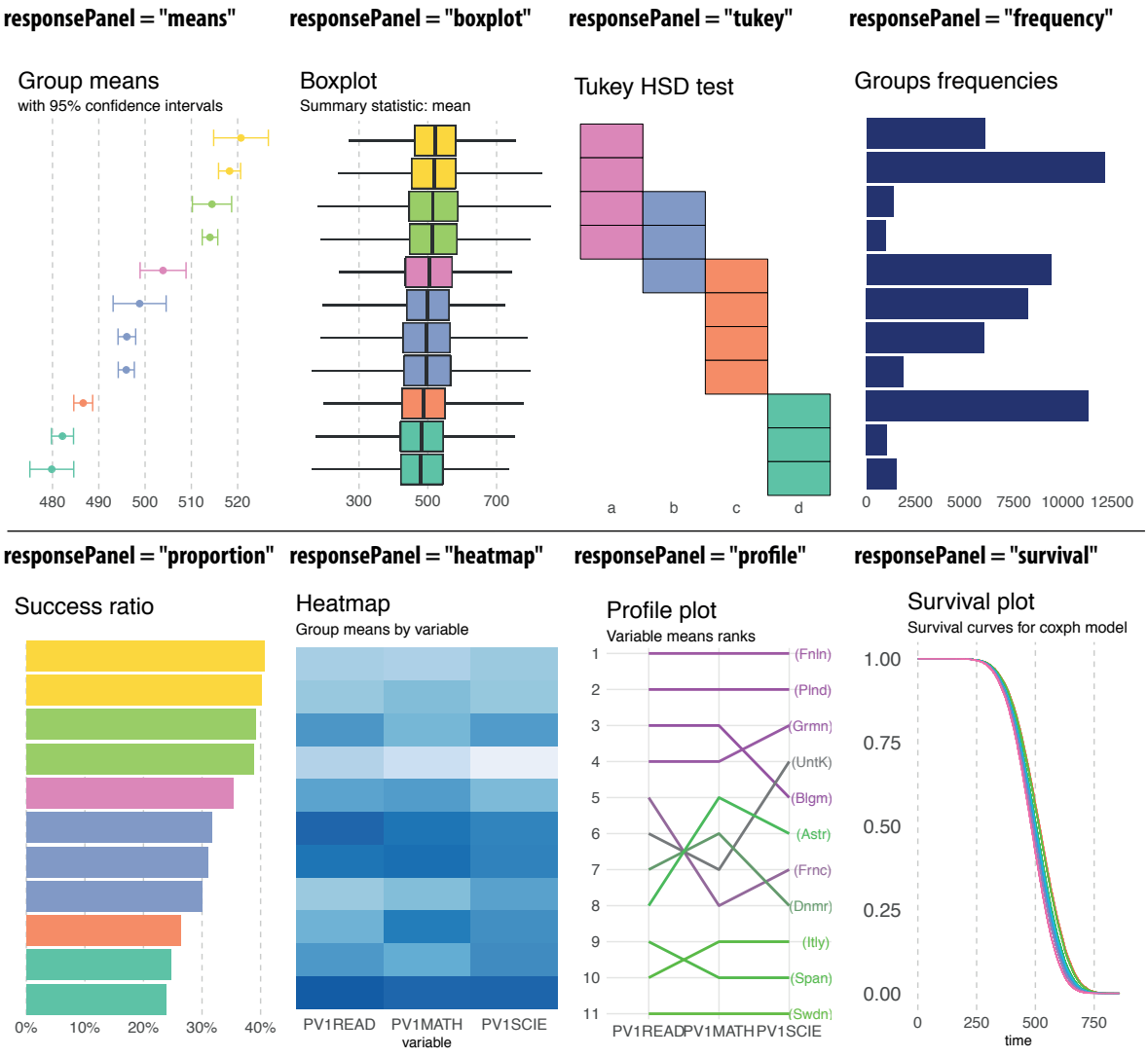


Figure 3.3: Avaliable options for the `responsePanel` argument. Different panels are designed to highlight different aspects of groups.

3.3. Optimal grouping selection

The Merge and Select algorithm returns a collection of models of different sizes / different numbers of groups. In order to select the best model first the optimization criterion must be specified. There are three metrics available in **factorMerger**:

- *Generalized Information Criterion* parameterized by the penalty. If this option is chosen, the model with the lowest GIC is returned.
- *p-value* for the Likelihood Ratio Test against the full model. If we go with this metric we choose the latest model in the merging path whose p-value for the LRT test against the full model is greater than a given threshold.
- *log-likelihood* of a model. A similar search is performed as in the previous point, but with models log-likelihood as the model score.

The most natural approach is to pick a model that minimizes the Generalized Information Criteria

$$GIC(\mathcal{M}) = -2l(\mathcal{M}) + p|\mathcal{M}|.$$

Here $|\mathcal{M}|$ denotes the number of groups in model, $\mathcal{M}$ while p is a penalty for model complexity. GIC corresponds to Akaike Information Criterion (AIC) for $p = 2$ or Bayesian Information Criterion (BIC) for $p = \log(n)$.

To ease the selection of the best model the bottom-left panel presents GIC scores for models in the merging path in the GIC plot. An example of such plot is presented in the Figure 3.4.

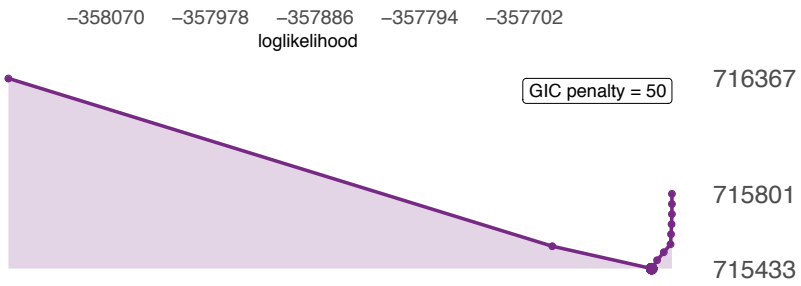


Figure 3.4: The GIC plot. OX axis corresponds to the log-likelihood for a model. OY axis corresponds to the GIC score for a model. Each dot denotes a single model from the merging path. GIC scores for the best, smallest and largest models are presented in the right axis.

3.4. The Fusing Strategy

The Algorithm 1 presents a general strategy for merging groups. The fully adaptive strategy is time consuming and may be slow for the large number of groups. Thus in the **factorMerger** package we have implemented four versions of the merging algorithms. These versions are summarized below.

Depending on the specific goal some steps of the Algorithm 1 may be performed differently. Possible options are:

- `method = "adaptive"`. The objective function is the logarithm of likelihood. The set *pairsSet* contains all possible pairs of groups available in a given step. Pairwise LRT distances are recalculated every step. This option is the slowest one since it requires the largest number of comparisons. It requires $O(k^3)$ model evaluations.

- `method = "fast-adaptive"`. Note that computing an objective function can be expensive and, especially for big datasets, it may be beneficial to limit the set of pairs that shall be compared. Also note that it is more likely that a pair of levels i and j will be chosen to merge if corresponding group averages are close. In this option, the objective function is the logarithm of likelihood, but the *pairsSet* is generated differently, in a following way. For Gaussian family of response, at the very beginning, the groups are ordered according to increasing averages and then *pairsSet* contains only pairs of closest groups. For other families the order corresponds to beta coefficients in a regression model. The detailed rules of ordering levels are given in the Table 3.2. This option is much faster than `method = "adaptive"` and requires $O(k^2)$ model evaluations.

- `method = "fixed"`. This option is based on the DMR algorithm introduced in Maj-Kańska et al. (2015). It was extended to cover survival models (however, for survival models there are no theorems of model selection consistency yet proven). The largest difference between this option and the `method = "adaptive"` is, that in the first step a pairwise distances are calculated between each groups based on the *LRT* statistic. Then the agglomerative clustering algorithm is used to merge consecutive pairs. It means that pairwise model differences are not recalculated as LRT statistics in every step but the `complete linkage` is used instead. This option is very fast and requires $O(k^2)$ comparisons.

- `method = "fast-fixed"`. This option may be considered as a modification of the `method = "fixed"`. Here, similarly as in the `"fast-adaptive"` version, we assume that if groups A, B and C are sorted according to their increasing beta coefficients, then it is worthwhile to join groups A and B or groups B and C (but not groups A and C). This assumption enables implementation of the `complete linkage` clustering to be more efficient, performed in a dynamic manner. The biggest difference is that in the first step we do not calculate the whole matrix of pairwise differences, but instead only differences between consecutive groups are measured. Then in each step only a single distance is calculated. This reduces the number of model evaluations to $O(k)$. Detailed description of beta coefficients is given in the Table 3.2.

family	**ordering statistic for a given group**
one-dimensional Gaussian	average in a group
multi-dimensional Gaussian	average in a group after the Kruskal's non-metric multidimensional scaling (Venables & Ripley 2002) to a one-dimensional space
binomial	proportion of successes in a group
survival	logarithm of a hazard ratio for a group

Table 3.2: Factor ordering by model family for `method = "fast-adaptive"` and `method = "fast-fixed"`

Described options differ in two ways. First, they differ in terms of computational time.

The fastest option is to preliminarily sort groups and then use the dynamic complete-linkage hierarchical algorithm which enables to join only adjacent groups. The slowest option is to calculate pairwise differences between groups after each fusion. Time performance comparisons are presented in Figure 3.5.

At the same time, the slowest option is the most accurate one, in terms that it gives models paths with highest log-likelihoods and is stable. A simple example that brings closer those characteristics is visualized in the Figure 3.6.

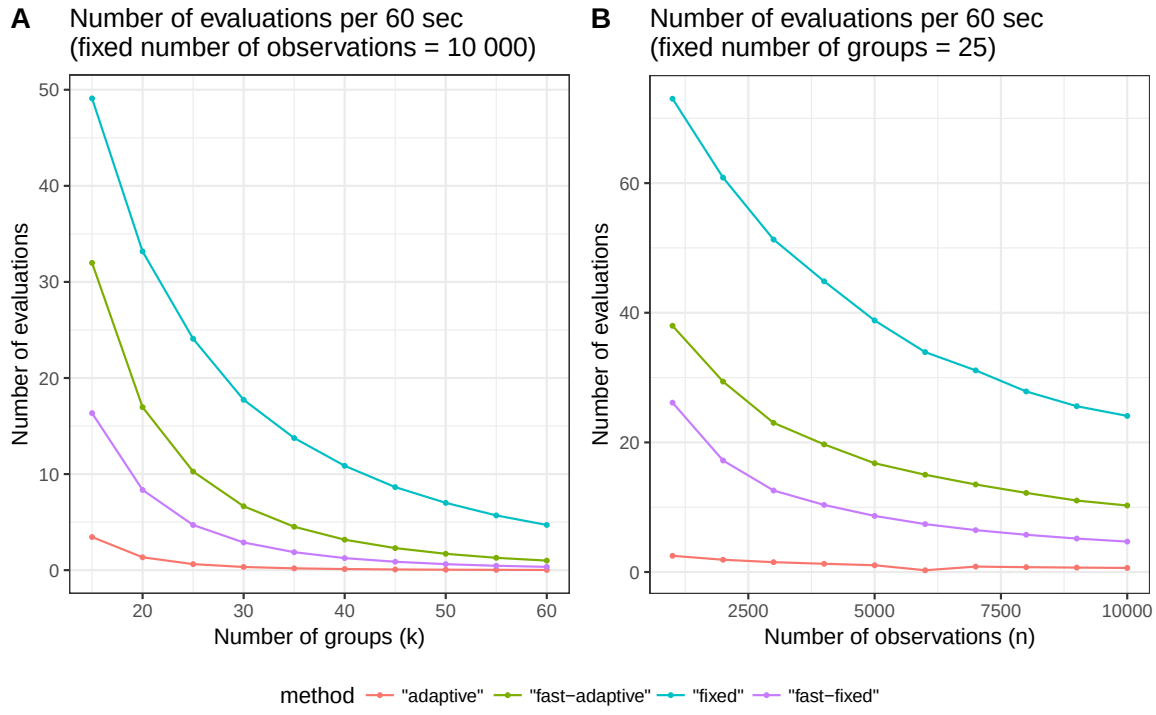


Figure 3.5: Average number of evaluations per 60 seconds for different number of groups (Panel A) and different sample sizes (Panel B). Fastest algorithms are those which limit comparisons only to consecutive groups (`"fast-adaptive"` and `"fast-fixed"`). For 10 observations methods `"fast-fixed"` and `"fast-adaptive"` are 5 times faster than method `"adaptive"`. For 60 observations those evaluation time ratios grow up to 200 and 42, respectively for `"fast-fixed"` and `"fast-adaptive"` against `"adaptive"`.

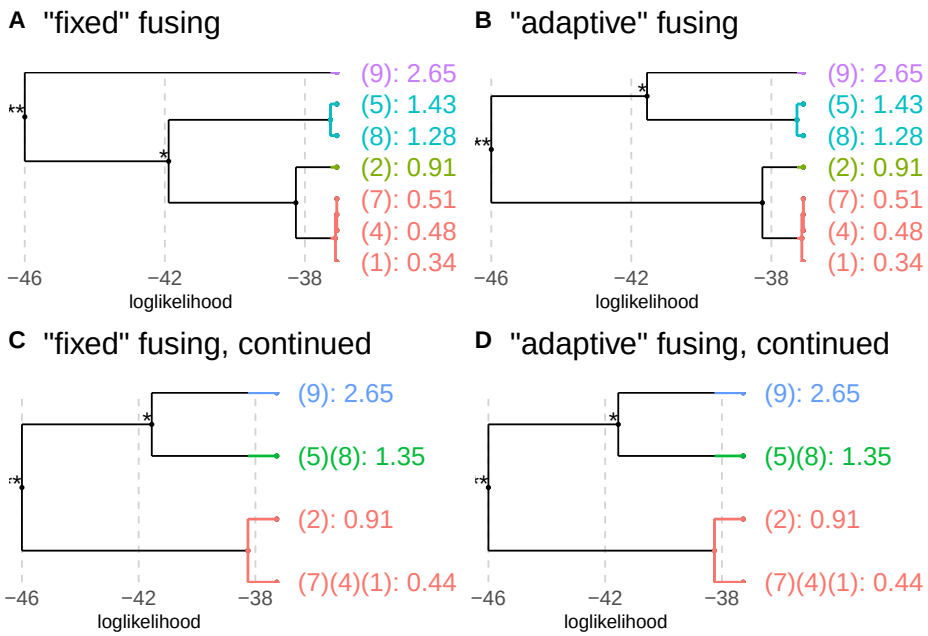


Figure 3.6: A comparison of methods `"fixed"` (left panels) and `"adaptive"` (right panels). We start with a sample consisting of 7 subgroups (top panels). First four steps of both algorithms are the same, but then the `"fixed"` algorithm chooses to merge groups (8)(5) and (9), while the `"adaptive"` algorithm goes with groups (1)(4)(7)(2) and (8)(5). The latter results in a model with higher log-likelihood (`"fixed"`: -41.89 vs. `"adaptive"`: -41.57). Note that if we choose different starting point (bottom panels) the `"fixed"` algorithm changes its path.

Chapter 4

Examples

The **factorMerger** package is highly customizable. In this chapter we present three different case studies to illustrate the use of **factorMerger** in real world examples. Each scenario is associated with a particular model family and also presents specific function arguments in action.

4.1. Gaussian models - what can you do with factorMerger?

PISA (*Program for International Students Assessment* 2012) is a program maintained by OECD to gather information on students' performance on various cognitive tests in countries around the world. Based on multiple survey results so-called *plausible values* are modeled which resemble individual test scores of students and can be used as a comparable measure. Those values are calculated in three different fields: mathematics, science and reading. Moreover, plausible values are normally distributed in the population and have conditionally Gaussian distribution.

The **factorMerger** package provides a student-level dataset (`data("pisa2012")`) which contains all tree plausible values, together with country affiliations of students. The data is a weighted version of the original data from the **PISA2012lite** package (Biecek 2015). A total number of rows is 271322 and a total number of groups is 43.

The model predicts plausible values in mathematics considering only single explanatory variable, country.

```
library(factorMerger)
library(dplyr)

data("pisa2012")

# to speed up the evaluation we use "fast-fixed" method
oneDimPisa <- mergeFactors(response = pisa2012$math,
                           factor = pisa2012$country,
                           method = "fast-fixed")
```

Note that it is only one command needed to perform the merging procedure.

We can use the obtained object to display the history of merging — each row of the table describes one step of the algorithm.

```
mergingHistory(oneDimPisa, showStats = TRUE) %>%
        head(5)
```

```
#   groupA groupB    model pvalVsFull pvalVsPrevious
# 0                -1606256     1.0000          1.0000
# 1 (Swdn) (SlvR) -1606256     0.9932          0.9932
# 2 (RssF) (Span) -1606256     0.9997          0.9805
# 3 (Chil) (Mlys) -1606256     0.9999          0.9459
# 4 (Frnc) (UntK) -1606256     1.0000          0.9213
```

Here, in the second step Russian Federation (RssF) and Spain (Span) were united. Log-likelihood, whose value is given in the model column, decreased marginally, p-values for the LRT test against the full model and against the previous model were 0.9997 and 0.9805, respectively. This means that the data partition created after two joins is equally good as the previous one and as the initial one.

Let us create a data partition based on the model with the lowest AIC in the merging path (GIC penalty = 2). Final grouping names are concatenations of original levels names.

```
aicPrediction <- cutTree(oneDimPisa,
                         stat = "GIC",
                         value = 2)

aicPrediction %>%
        table() %>%
        head(4)
```

```
#  (Clmb)  (Brzl) (Mntn)(Urgy) (Chil)(Mlys)
#    8902   38525           763         10872
```

We can also see the final data partition in a table. Below original group labels are printed in the abbreviated form (orig) together with their final cluster name (pred). For example, Poland (Plnd) ends up in the (Cand) (Plnd) group, which consists of two members: Poland and Canada (Cand).

```
getOptimalPartitionDf(oneDimPisa,
                      stat = "GIC",
                      value = 2) %>%
    head(5)
```

```
#     orig          pred
# 1 (RssF) (RssF)(Span)
# 2 (Blgm) (Grmn)(Blgm)
# 3 (Grmn) (Grmn)(Blgm)
# 4 (Kore)        (Kore)
# 5 (Plnd) (Cand)(Plnd)
```

4.2. The binomial model - GIC manipulation

This chapter uses a dataset included in the **factorMerger** package (`data("ess")`) on happiness of 21 European countries based on the European Social Survey (*European Social Survey* 2014). A binary variable called *happy* specifies if a given individual considers themselves a happy person (or, more precisely, his/her answer for the question "Taking all things together, how happy would you say you are?" was greater than 5 in a scale of 0 to 10). Again, data is weighted according to original weights given by ESS. A total number of rows in `ess` is 200075, there are 21 countries included.

By default the `plot.factorMerger` function uses GIC with the penalty equal to two (Akaike Information Criterion). However, in some cases it may not be restrictive enough. Since the number of observation is large, we may use large GIC penalty, like 500.

```
library(factorMerger)
data("ess")

happyMerge <- mergeFactors(ess$happy, ess$country,
                           family = "binomial",
                           method = "fast-fixed")

# GIC as AIC
p1 <- plot(happyMerge, panel = "GIC",
           title = "",
           panelGrid = FALSE)

# GIC as BIC
p2 <- plot(happyMerge, panel = "GIC",
           penalty = log(NROW(ess$country)),
           title = "",
           panelGrid = FALSE)

# Large GIC
p3 <- plot(happyMerge, panel = "GIC",
           penalty = 500,
           title = "",
           panelGrid = FALSE)
```

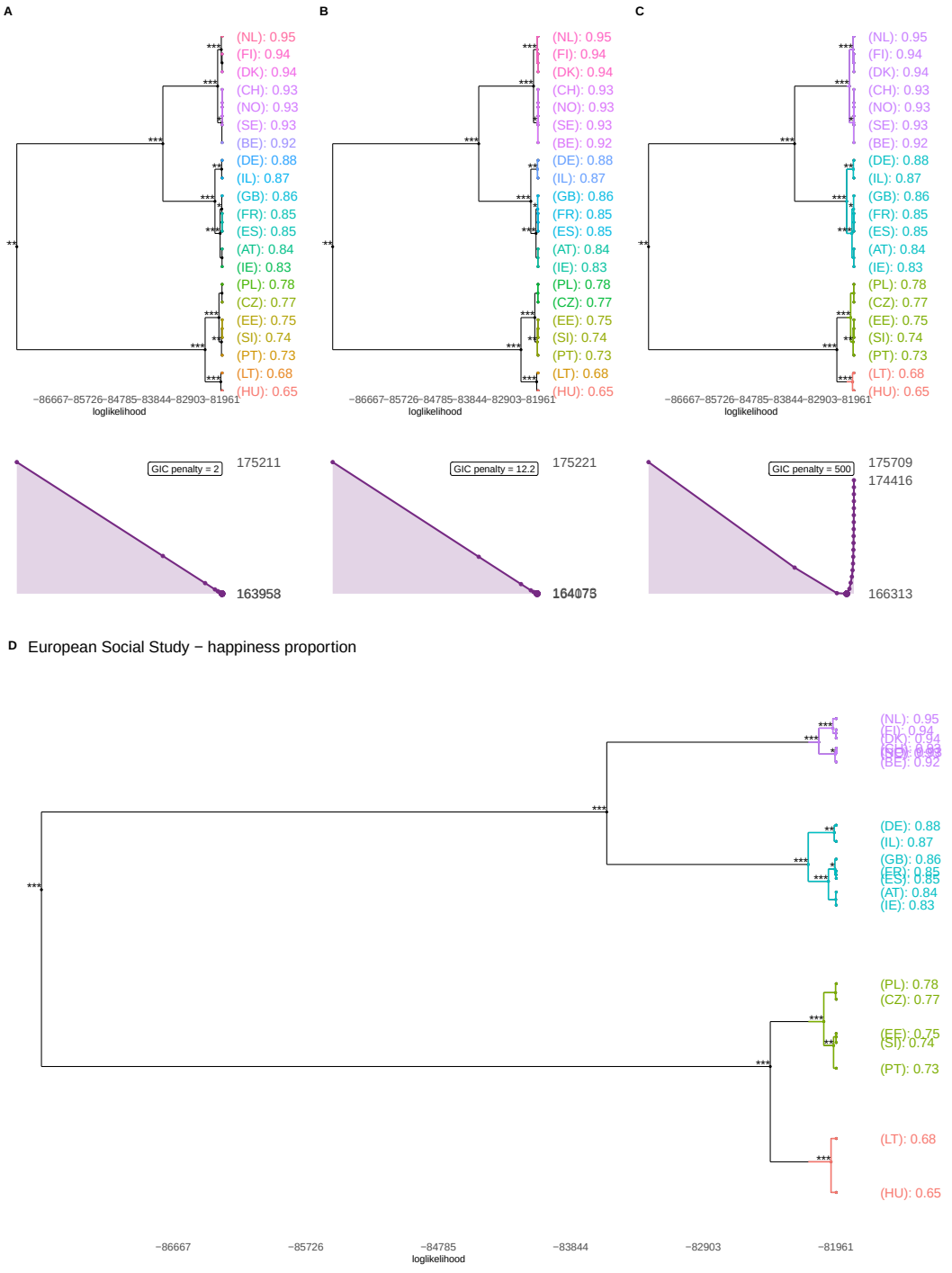


Figure 4.1: Results for GIC with different penalties: AIC with penalty = 2 (Panel A), BIC with penalty = 12.2 (Panel B) and GIC with penalty = 500. Optimal numbers of groups for those penalties are: 17, 9, 4. The Merging Path Plot for GIC with penalty = 500 is presented in Panel D. Nodes positions on OY axis correspond to fractions of happy citizens in given country / groups of countries.

4.3. The survival model - customization of the visualization

In this example we use data from the The Cancer Genome Atlas Project (*The Cancer Genome Atlas Wiki* 2015) from the **RTCGA.clinical** package (Kosinski 2016). TCGA is a public funded project that aims to catalogue and discover major cancer-causing genomic mutations to create a comprehensive *atlas* of cancer profiles. The **RTCGA.clinical** package provides a snapshot of those clinical data created at 2015-11-01. In our example we focus on patients who suffer from breast cancer and are treated with different drugs. We are interested whether drug treatments may be grouped according to their effectiveness.

The dataset used in this example is included in **factorMerger** (`data("BRCA")`).

First, some data preprocessing is performed.

```
library(factorMerger)
library(dplyr)
library(forcats) # fct_lump
library(survival) # Surv

data("BRCA")

BRCA <- BRCA %>%
    filter(!is.na(drugName))

# use only highly represented groups
drugName <- fct_lump(BRCA$drugName, prop = 0.05)
brcaSurv <- Surv(time = BRCA$time,
                 event = BRCA$vitalStatus)

drugMerge <- mergeFactors(response = brcaSurv,
                          factor = drugName,
                          family = "survival",
                          method = "adaptive")
```

Now we can plot the result. By default four panels are included in the plot, the tree is colored to denote final clusters, nodes are spaced using equal distance.

Let's add some customization!

```
plot(drugMerge,
     # show groups' effects in the tree
     nodesSpacing = "effects",
     # add custom title
     title = "BRCA: patient survival vs. drug treatment",
     # display only two panels
     panel = "response",
     # do not color clusters
     colorClusters = FALSE,
     # mark optimal model with a vertical line
     showSplit = TRUE,
     # set your favorite RColorBrewer palette
     palette = "Dark2")
```

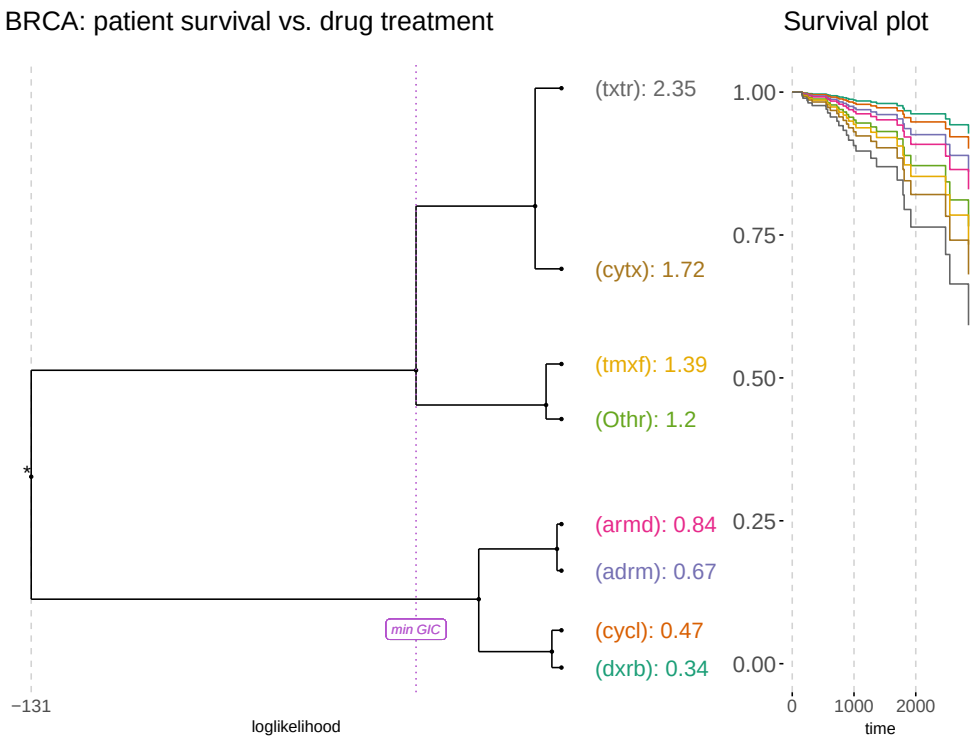


Figure 4.2: Customized `plot.factorMerger` output. Colors of OY axis' labels are guides for the right panel. Tree nodes are spaced according to group effects. A vertical line is added to mark the optimal data partition.

Chapter 5

Summary and Future Directions

Merge and Select is a novel approach to summarize groups similarities based on the LRT statistic. It is a useful tool to explore group similarities in k-sample problems.

In this thesis we have presented the methodology and its applications of this methodology as it is implemented in the **factorMerger** R package. Currently the implementation is limited to a single grouping variable. The natural extension of this approach is related to including more grouping variables, possibly with interactions. Other possible updates are related to different classes of models. Instead of the Likelihood Ratio Test other tests may be used. For example, the Wilcoxon test may be used for semi-parametric modeling.

Even with these limitations **factorMerger** is a useful tool for exploratory analysis as it helps to summarize structure of groups' in a single plot.

Chapter 6

Acknowledgements

We acknowledge the financial support from the *NCN Opus grant 2016/21/B/ST6/02176*.

Bibliography

Biecek, P. (2015), 'PISA2012lite: ready PISA2012'.
URL: *https://doi.org/10.5281/zenodo.17866*

Bondell, H. D. & Reich, B. J. (2008), 'Simultaneous factor selection and collapsing levels in ANOVA', *Department of Statistics, North Carolina State University* .

Cox, D. R. (1992), Regression models and life-tables, *in* 'Breakthroughs in statistics', Springer, pp. 527–541.

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

European Social Survey (2014).
URL: *http://www.europeansocialsurvey.org*

Fox, J. & Weisberg, S. (2011), *An R Companion to Applied Regression*, second edn, Sage, Thousand Oaks CA.
URL: *http://socserv.socsci.mcmaster.ca/jfox/Books/Companion*

Hothorn, T., Bretz, F. & Westfall, P. (2008), 'Simultaneous inference in general parametric models', *Biometrical Journal* **50**(3), 346–363.

Kosinski, M. (2016), *RTCGA.clinical: Clinical datasets from The Cancer Genome Atlas Project*. R package version 20151101.6.0.

Lenth, R. V. (2016), 'Least-squares means: The R package lsmeans', *Journal of Statistical Software* **69**(1), 1–33.

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

Maj-Kańska, A., Pokarowski, P., Prochenka, A. et al. (2015), 'Delete or merge regressors for linear model selection', *Electronic Journal of Statistics* **9**(2), 1749–1778.

Peter Rousseeuw & Leonard Kaufman (1990), *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley, New York.

Program for International Students Assessment (2012).
URL: *http://www.oecd.org/pisa/pisaproducts/pisa2012database-downloadabledata.htm*

R Core Team (2017), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://www.R-project.org/*

The Cancer Genome Atlas Wiki (2015).
URL: *https://wiki.nci.nih.gov/display/TCGA*

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. & Knight, K. (2005), 'Sparsity and smoothness via the fused lasso', *Journal of the Royal Statistical Society* pp. 01–108.

Tukey, J. (1949), 'Comparing Individual Means in the Analysis of Variance', *BIOMETRICS* pp. 99–114.

Venables, W. N. & Ripley, B. D. (2002), *Modern Applied Statistics with S*, fourth edn, Springer, New York. ISBN 0-387-95457-0.
URL: *http://www.stats.ox.ac.uk/pub/MASS4*

Wickham, H. (2009), *ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag New York.
URL: *http://ggplot2.org*

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