Simultaneous Deleting or Merging Regressors for Linear Model Selection

Aleksandra Maj[1], Agnieszka Prochenka[1], and Piotr Pokarowski[2]

[1] Institute of Computer Science, Polish Academy of Sciences,
ul. Jana Kazimierza 5, 01-237 Warsaw, Poland
[2] Institute of Applied Mathematics and Mechanics, University of Warsaw,
ul. Banacha 2, 02-097 Warsaw, Poland

Abstract

In the article we consider a problem of simultaneous deleting continuous variables and merging levels of factors in linear model. We propose a backward selection procedure called DMR in two variants: the first is similar to the backward stepwise regression and the second, faster implementation combines the agglomerative clustering of levels of factors with ranking regressors by squared t-statistics. In the paper we show that our algorithm is consistent. For the formulated problem we also propose a generalization of performance measures such as sensitivity and specificity. We present a simulation study, which shows substantial advantage of DMR over other methods described in the literature.

Keywords: ANOVA; Bayes Factors; GIC; Hierarchical clustering; Merging levels; QR decomposition; Sensitivity; Similarity; Specificity; T-statistic; Variable selection.

1 Introduction

Model selection is usually understood as selection of explanatory variables. However, when a categorical predictor is considered, in order to reduce model's complexity, we can either exclude the whole factor or merge its levels.

A traditional method to examine the relationship between a continuous response and categorical variables is analysis of variance (ANOVA). However, ANOVA answers only a question of the overall importance of a factor. The next step of the analysis are pairwise comparisons of group means within important factors. Typically post-hoc analysis such as Tukey's honestly significant difference (HSD) test or multiple comparison adjustments (Bonferroni, Scheffe, Hochberg) is used. A drawback of pairwise comparisons is non-transitivity of conclusions.

As a motivating example, let us consider data `Cars93` from `R` library `MASS`. The relationship between logarithm of fuel consumption and other characteristics of 81 cars is modeled. The dependence between the response and the number of cylinders examined with the use of Tukey's HSD analysis (Figure 1) gives inconclusive answers: $\beta_4 = \beta_5$, $\beta_5 = \beta_6$, but $\beta_4 \neq \beta_6$.

In the article we introduce a novel procedure called delete or merge regressors (DMR), which enables efficient search among partitions of factor levels, hence the issue of non-transitivity does not occur. When applying DMR procedure to the `Cars93` data, the number of parameters is efficiently diminished from 31 to 11 with no considerable loss in R^2 (from 0.92 to 0.9), while a model received from the stepwise backward model selection minimizing BIC implemented in the `stepAIC` function in `R` has 14 parameters with $R^2 = 0.89$.

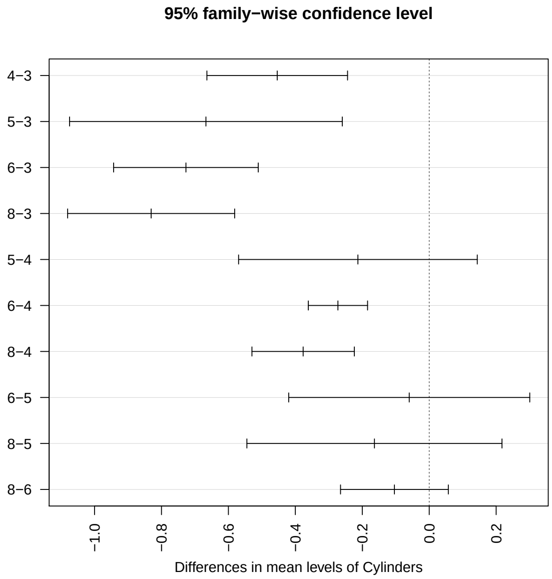


Figure 1: Results of Tukey's HSD.

The idea of partitioning a set of levels of a factor into non-overlapping subsets has already been discussed in the literature. In the article [1], 1949, Tukey proposed a stepwise backward procedure based on the Studentized range methods. Other methods performing multiple comparison procedures, based on clustering means in ANOVA, were described by Scott and Knott, 1974 [2], Calinski and Corsten, 1985 [3] and Corsten and Denis, 1990 [4]. However, these methods do not generalize directly to the problem with any number of factors.

Also the problem of simultaneous continuous variables selection and merging levels of factors is present in the literature. A method introduced by Bondell and Reich, 2009, [5] called collapsing and shrinkage ANOVA (CAS-ANOVA) solves the problem with the use of the least absolute shrinkage and selection operator (LASSO; Tibshirani, 1996, [6]), where the L_1 penalty is imposed on differences between parameters corresponding to levels of each factor. Gertheiss and Tutz, 2011, [7] proposed a modification of CAS-ANOVA, which is more computationally efficient because of using the least angle regression (LARS; Efron et al., 2004, [8]) algorithm.

We propose a backward selection procedure called delete or merge regressors

(DMR), which combines deleting the continuous variables with merging levels of factors. The method assumes greedy search among linear models with a set of constraints of two types: either a parameter for a continuous variable is set to zero or parameters corresponding to two levels of a factor are set to equal each other. DMR is a stepwise regression procedure, where in each step a new constraint is added according to ranking of the hypotheses based on squared t-statistics. As a result a nested family of linear models is obtained and the final decision is made according to minimization of generalized information criterion (GIC).

Two variants of DMR are described in the article. The first, more greedy version adapts agglomerative clustering, where squared t-statistics define the dissimilarity measure. This procedure generalizes concepts introduced by Ciampi et al., 2008, [9] and Zheng and Loh, 1995, [10]. The second version assumes recalculation of t-statistics in each step, which causes loss in computational efficiency.

In the paper we show that DMR algorithm is consistent. The time complexity of the more greedy version of DMR is $O(np^2)$, where n is the number of observations and p is the number of parameters in the full model. We describe also a simulation study and discuss a pertaining `R` package. The simulations show that DMR is several hundred times faster with significantly lower error of selection than CAS-ANOVA.

The next problem considered in the article is determining the quality of performance of model selection. Commonly used measures are for example true positive rate (TPR) or false negative rate (FNR). In the literature [5], [7] a generalization of these rates to the problem of partitioning factor levels can be found. However, these measures tend to diminish the influence of continuous predictors and factors with a small number of levels. We propose a different generalization which is based on the dimension of linear subspace of the parameter space defined by the imposed constraints.

The remainder of the article proceeds as follows. The class of feasible models considered when performing model selection is defined in Section 2. DMR procedures are introduced in Section 3, while the asymptotic properties are discussed in Section 4. Generalization of measures of performance is introduced in Section 5. Simulations and real data example are given in Section 6 to illustrate the method. All proofs are given in the Appendix.

2 Feasible models

Let us consider a full rank linear model with n observations and $p < n$ parameters:

$$y = X\beta + \varepsilon = \mathbf{1}_n\beta_{00} + X_0\beta_0 + X_1\beta_1 + \ldots + X_l\beta_l + \varepsilon, \tag{1}$$

where:

1. $\varepsilon \sim \mathcal{N}(0, \sigma^2\mathbb{I}_n)$.
2. $X = [\mathbf{1}_n, X_0, X_1, \ldots, X_l]$ is the model matrix divided as follows: X_0 is the matrix corresponding to continuous regressors and $X_1, \ldots, X_l$ are the zero-one matrices encoding corresponding factors with the first level set as reference.

3. $\beta = [\beta_{00}, \beta_0^T, \beta_1^T, \ldots, \beta_l^T]^T$ is the parameter vector divided as follows: β_{00} is the intercept, $\beta_0 = [\beta_{10}, \ldots, \beta_{p_0 0}]^T$ is a vector of coefficients for continuous variables and $\beta_k = [\beta_{2k}, \ldots, \beta_{p_k k}]^T$ is a vector of parameters corresponding to the k-th factor, $k = 1, \ldots, l$, hence the length of parameter vector is $p = 1 + p_0 + (p_1 - 1) + \ldots + (p_l - 1)$.

Definition 1. *An* elementary hypothesis *for linear model (1) is a linear hypothesis of one of two types:*

(*) $h_{jk} : \beta_{jk} = 0$ *for all* j, k *or*
(**) $h_{ijk} : \beta_{ik} = \beta_{jk}$ *for all* i, j *and* $k > 0$.

Definition 2. *A* feasible model *is defined as a sequence* $m = (C, P_1, ..., P_l)$*, where* C *denotes a subset of continuous variables and* P_k *is a partition of levels of the* k*-th factor. Such a model can be encoded by a set of elementary hypotheses. A set of all feasible models is denoted by* $\mathcal{M}$*.*

2.1 Change of variables

In order to replace a constrained by an unconstrained optimization problem a change of variables in model m is performed. The model can be defined by a set of following equations

$$\begin{cases} y = X\beta + \varepsilon \\ A_C\beta = 0, \end{cases}$$

where A_C is a matrix of elementary hypotheses describing constraints induced by the model.

Let us define a square matrix

$$A = \begin{bmatrix} A_1 \\ A_0 \end{bmatrix},$$

where A_0 is a matrix of elementary hypotheses in a convenient form described below, with rows spanning the same space as rows of A_C and A_1 is a complement of A_0 to a square matrix. We want A_0 and A_1 to satisfy

$$A = \begin{bmatrix} A_1 \\ A_0 \end{bmatrix} = \begin{bmatrix} \mathbb{I}_q & 0 \\ B & \mathbb{I}_{p-q} \end{bmatrix}.$$

Matrix A of such a form uniquely encodes model m and can always be obtained by appropriate permutation of the columns of the model matrix, which can be performed in the following way:

1. For each factor with partition P_k, $k = 1, \ldots, l$ with i_k clusters, where $P_k = \{U_1^k, \ldots, U_{i_k}^k\}$, rename its levels so that

$$U_j^k = \{j, i_k + \sum_{s=1}^{j-1} |U_s^k| - j + 2, \ldots, i_k + \sum_{s=1}^{j} |U_s^k| - j\},$$

for $j = 1, \ldots, i_k$.

2. Sort the columns of the model matrix in the following order:
   (a) intercept,
   (b) all continuous variables present in the model (these for which the beta coefficient is non-zero),
   (c) i_k first levels of k-th factor, $k = 1, \ldots, l$,
   (d) all remaining continuous variables,
   (e) all remaining levels of factors.

Further in the article we assume that the columns of model matrices for considered models are permuted so that the constraints matrices have a form such as A_0.

Example 1. As an illustrative example consider a model consisting of one factor with $P_1 = \{\{1,2,3\},\{4,5\},\{6,7,8\}\}$. This parametrization corresponds to the following constraint matrix:

$$A_C = \left[\begin{array}{cccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ \hline 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 \end{array}\right].$$

After renaming levels in the way described above, we get partition P_1 of the form $P_1 = \{\{1,4,5\},\{2,6\},\{3,7,8\}\}$ and matrix

$$A = \left[\begin{array}{c} A_1 \\ \hdashline A_0 \end{array}\right] = \left[\begin{array}{cccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ \hline 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ \hdashline 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 0 & 1 \end{array}\right]$$

in the desired form.

Thanks to such reparametrization of the model, transition from constrained to unconstrained problem is immediate.

$$X\beta = XA^{-1}A\beta = Z\xi,$$

where $Z = XA^{-1}$ and $\xi = A\beta$. From Schur complement we get

$$A^{-1} = \begin{bmatrix} \mathbb{I}_q & 0 \\ -B & \mathbb{I}_{p-q} \end{bmatrix} = [A^1, A^0],$$

since $A_0\beta = 0$ we have

$$Z\xi = X[A^1, A^0]\begin{bmatrix} \xi_1 \\ 0 \end{bmatrix} = Z_1\xi_1,$$

where $Z_1 = XA^1$, $\xi_1 = A_1\beta$. The linear space of parameters changes from $\mathcal{L} = \{\beta \in \mathbb{R}^p : A_0\beta = 0\}$ to $\mathcal{L}(A^1) = \{A^1\xi_1 : \xi_1 \in \mathbb{R}^q\}$. The dimension of space $\mathcal{L}(A^1)$ is called the size of model m and denoted by $|m|$.

Note that imposing an elementary hypothesis on parameter vector β is equivalent, in terms of the new parametrization, to either eliminating one column or summing up two columns of the model matrix, which decreases its number of columns by one. These operations are explicitly visible from the form of A^1.

Definition 3. *We define* the inclusion relation between two models m_1 and m_2 *by inclusion of linear spaces spanned by columns of the corresponding matrices* $A^1_{m_1}$ *and* $A^1_{m_2}$*:*

$$m_1 \subseteq m_2 \iff \mathcal{L}(A^1_{m_1}) \subseteq \mathcal{L}(A^1_{m_2}).$$

2.2 Generalized Information Criterion

Definition 4. Generalized Information Criterion *for model* m *is defined as:*

$$GIC(m) = n \log\left(RSS_m\right) + r_n(|m| + 1),$$

where r_n *is the penalty for model size.*

The goal of our method is to find the best feasible model according to GIC, taking into account that the number of feasible models grows exponentially with p. Since for the k-th factor number of possible partitions is the Bell number $\mathcal{B}(p_k)$, the number of all feasible models is $2^{p_0}\prod_{k=1}^{l}\mathcal{B}(p_k)$. In order to significantly reduce the amount of computations, we propose the following greedy backward search.

3 Algorithms

Assuming that X is of full rank the QR decomposition of the model matrix is

$$X = Q_pR,$$

where Q_p is $n \times p$ orthogonal matrix and R is $p \times p$ upper triangular matrix. Then

$$z = Q_p^Ty, \tag{2}$$

$$\widehat{\sigma}^2 = \frac{\|(I - Q_pQ_p^T)y\|^2}{n-p} \tag{3}$$

and

$$\widehat{\beta} = R^{-1}z.$$

Let us define set of indexes corresponding to continuous variables and factors

$$\mathrm{Ind}_0 = \{0, 1, \ldots, p_0\},\ \mathrm{Ind}_k = \{2, \ldots, p_k\} \text{ for } k = 1, \ldots, l.$$

Then

$$\widehat{\beta} = (\widehat{\beta}_{jk})_{j\in\mathrm{Ind}_k} = (r_{jk}^Tz)_{j\in\mathrm{Ind}_k} \text{ for } k = 0, \ldots, l,$$

where

$$R^{-1} = [r_{00}^T, r_{10}^T, \dots, r_{p_0 0}^T, r_{21}^T, \dots, r_{p_1 1}^T, \dots, r_{2l}^T, \dots, r_{p_l l}^T]^T. \tag{4}$$

Elementary hypotheses of types (*) and (**) defined in Definition 1 can be now rewritten as:

(*) h_{jk}: $\beta_{jk} = 0 \iff a_{1jk}^T \beta = 0$, where $a_{1jk} = a_{st}(j,k) = \mathbb{1}(s = j, t = k)$, where $s \in \mathrm{Ind}_t$ and $t = 0, \dots, l$,

(**) h_{ijk}: $\beta_{ik} = \beta_{jk} \iff a_{ijk}^T \beta = 0$, where $a_{ijk} = a_{st}(i,j,k) =$ $= \mathbb{1}(s = i, t = k) - \mathbb{1}(s = j, t = k)$, where $s \in \mathrm{Ind}_t$ and $t = 1, \dots, l$.

3.1 DMR algorithm

1. Perform the QR decomposition of the full model matrix, getting matrix R^{-1}, vector z and variance estimator $\widehat{\sigma}^2$ as in equations (4), (2) and (3).
2. Calculate squared t-statistics:
   (a) for all elementary hypotheses of type (*):
   $$T_{1jk}^2 = \frac{\widehat{\beta}_{jk}^2}{\widehat{Var}(\widehat{\beta}_{jk})} = \frac{(r_{jk}^T z)^2}{\widehat{\sigma}^2 \|r_{jk}\|^2} \quad \text{for} \quad k \geq 0,\ j \in \mathrm{Ind}_k \setminus \{0\},$$
   (b) for all elementary hypotheses of type (**):
   $$T_{ijk}^2 = \frac{(\widehat{\beta}_{ik} - \widehat{\beta}_{jk})^2}{\widehat{Var}(\widehat{\beta}_{ik} - \widehat{\beta}_{jk})} = \frac{((r_{ik} - r_{jk})^T z)^2}{\widehat{\sigma}^2 \|r_{ik} - r_{jk}\|^2}$$
   for $k > 0$, $i, j \in \mathrm{Ind}_k$.
3. For each $k > 0$ perform agglomerative clustering using $D_k = [d_{ijk}]_{ij}$ as dissimilarity matrix, where:
   (a) $d_{1jk} = d_{i1k} = T_{1jk}^2$ for $i, j \in \mathrm{Ind}_k$,
   (b) $d_{ijk} = T_{ijk}^2$ for $i, j \in \mathrm{Ind}_k$, $i \neq j$,
   getting vectors of cutting heights g_k, $k = 1, \dots, l$.
4. Combine all vectors g_k with g_0 defined as $g_0 = [T_{110}^2, \dots, T_{1p_0 0}^2]$, denote the given vector as g. Sort g in increasing order: $g = [g^{(1)}, \dots, g^{(p-1)}]^T$. Every element $g^{(i)}$ corresponds to an elementary hypothesis a_i. A sequence of nested linear constraints on model parameters $A_i \beta = 0$ is obtained, where $A_i = [a_1, \dots, a_i]$, $i = 0, \dots, p-1$ and $A_0 = 0$.
5. Perform QR decomposition of matrix $R^{-T} A_{p-1}^T$ getting the orthogonal matrix $W = [w_1, \dots, w_{p-1}]$.
6. Set $\mathrm{RSS}_0 = \|y\|^2 - \|z\|^2$ and $\mathrm{GIC}_0 = n \log \mathrm{RSS}_0 + (p+1) r_n$ for model without constraints.
   For $i = 1, \dots, p-1$
   $$\mathrm{RSS}_i = \mathrm{RSS}_{i-1} + (w_i^T z)^2,$$
   where calculations are described in the Appendix A and
   $$\mathrm{GIC}_i = n \log \mathrm{RSS}_i + (p - i + 1) r_n.$$

7. Selected model $\widehat{m}$ is a model with hypotheses $A_{\widehat{i}}$ accepted, where

$$\widehat{i} = \operatorname*{arg\,min}_{0 \leq i \leq p-1} GIC_i.$$

The dominating operation in the described procedure is the QR decomposition of the full model matrix. Hence, the time complexity of DMR algorithm is $O(np^2)$.

This procedure assumes transitivity of hypotheses, for example accepting hypotheses $\beta_{ik} = \beta_{jk}$ and $\beta_{jk} = \beta_{mk}$ causes acceptance of hypothesis $\beta_{ik} = \beta_{mk}$. Notice that considered hypotheses can be ambiguously encoded. In order to avoid this problem the following convention will be used while cluster merging: U_1, $U_2 \subseteq \{1, \ldots, p_k\}$ clusters to merge, $U_1 \cap U_2 = \emptyset$, then $i_1 = \min_{i \in U_1} i$, $i_2 = \min_{i \in U_2} i$ and the hypothesis to accept is $\beta_{i_1 k} = \beta_{i_2 k}$.

An exemplary run of DMR algorithm is shown in Figure 2. The agglomerative clustering was performed for data, which is described in Section 6.1, consisting of three categorical variables. The horizontal dotted line indicates the cutting height for the best model chosen by BIC (special case of GIC, where $r_n = \log n$).

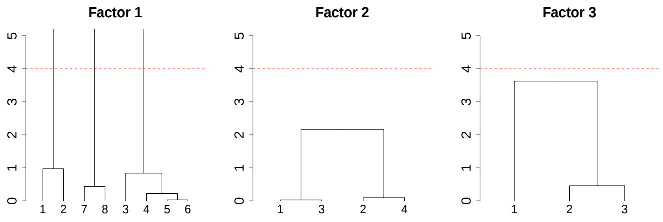


Figure 2: Dendrograms for exemplary run of DMR algorithm.

3.2 stepDMR algorithm

The stepDMR algorithm is based on the RSS calculation described in Appendix A, where the set of hypotheses under consideration is of the form $A_0\beta = 0$, hence the vector e is zero and equation (9) reduces to:

$$\|y - X\widehat{\beta}_c\|^2 = \sum_{i=1}^{p-q} (w_i^T z)^2. \tag{5}$$

By $S^{(i-1)}$ and $W^{(i-1)}$ we denote respectively the matrix corresponding to the set of possible hypotheses and the orthogonal matrix corresponding to the set of accepted hypotheses in the i-th step of the algorithm. Steps of the algorithm:

1. $S^{(0)} = R^{-T} A_{\text{all}}^T$, where A_{all} is the matrix of all possible elementary hypotheses.
2. $W^{(0)} = 0$. Calculate RSS for model without constraints getting RSS_0 and $\text{GIC}_0 = n \log \text{RSS}_0 + (p+1) r_n$.

3. Perform the orthogonal projection of possible elementary hypotheses $S^{(i-1)}$ onto the orthogonal complement of subspace spanned by already accepted hypotheses $W^{(i-1)}$:
$$V^{(i)} = (I - W^{(i-1)}(W^{(i-1)})^T)S^{(i-1)}.$$
4. Normalize columns of matrix $V^{(i)}$
$$w_j^{(i)} = \frac{v_j^{(i)}}{\|v_j^{(i)}\|}.$$
5. According to equation (9) calculate the vector of increases in residual sum of squares and choose the hypothesis corresponding to the minimal value.
$$\widehat{j} = \arg\min_j((w_j^{(i)})^T z)^2,$$
bind the chosen hypothesis with the matrix of already accepted hypotheses:
$$W^{(i)} = [W^{(i-1)}, w_{\widehat{j}}^{(i)}].$$
6. Calculate residual sum of squares
$$\mathrm{RSS}_i = \mathrm{RSS}_{i-1} + ((w_{\widehat{j}}^{(i)})^T z)^2$$
and GIC
$$\mathrm{GIC}_i = n \log \mathrm{RSS}_i + (p - i + 1)r_n.$$
7. Remove from $S^{(i-1)}$ columns linearly dependent with columns of $W^{(i)}$, getting new matrix of possible hypotheses $S^{(i)}$.
8. Go back to step 3 until the model is reduced to the intercept.
9. Selected model $\widehat{m}$ is a model with hypotheses corresponding to the columns of $W^{(\widehat{i})}$ accepted, where
$$\widehat{i} = \underset{0 \le i \le p-1}{\arg\min}\, GIC_i.$$

The dominating operation in each of p steps of the described procedure is the QR decomposition of the model matrix. Hence, the time complexity of stepDMR algorithm is $O(np^3)$.

4 Asymptotic properties of DMR algorithm

We first introduce some notations. In this section we use a simplifying notation $f_n \prec g_n$ which corresponds to $f_n = o(g_n)$. We allow the number of predictors p_n to grow with the number of observations n under the condition $p_n \prec n$.

We distinguish the following subsets of the set of all feasible models $\mathcal{M}$:

1. Full model f, which is a model without constraints.

2. Uniquely defined model t, which is minimal among true models in the sense of inclusion defined in Definition 3 and is fixed and does not depend on sample size. We assume that the model consists of a finite number of continuous variables and a finite number of factors with finite numbers of levels.
3. A set $\mathcal{M}_{\mathcal{V}}$ of models with one false elementary hypothesis accepted:

$$\mathcal{M}_{\mathcal{V}} = \{m \subseteq f : |m| = |f| - 1 \text{ and } t \not\subseteq m\},$$

4. A set $\mathcal{M}_{\mathcal{T}}$ of models with one true hypothesis accepted:

$$\mathcal{M}_{\mathcal{T}} = \{m \subseteq f : |m| = |f| - 1 \text{ and } t \subseteq m\}.$$

Theorem 1. *Let us denote*

$$d_n = \min_{m \in \mathcal{M}_{\mathcal{V}}} \beta_t^T X_t^T (\mathbb{I} - H_m) X_t \beta_t,$$

where X_t is the model matrix of the true model t with appropriate columns of the full model matrix X deleted or merged, $\beta_t \in \mathbb{R}^{|t|}$ is the parameter vector of t and H_m is the hat matrix corresponding to model m. Assuming that X is of full rank p_n, where $p_n \prec r_n \prec n$ and $p_n \prec d_n$ we have

$$\lim_{n \to \infty} \mathbb{P}(\widehat{m} = t) = 1,$$

where $\widehat{m}$ is the model selected by DMR procedure from Section 3.1, where the linkage criterion for hierarchical clustering is a convex combination of minimum and maximum of the pairwise distances between clusters.

Proof can be found in the Appendix B.

5 Measures of performance and quality of selection

5.1 Measures of performance

When performing simulations a researcher usually faces a problem of comparing results with the underlying truth. Furthermore, one would like to have a measure of performance which is more liberal than a binary response, whether the true model was correctly identified or not. Traditionally for model selection with only continuous predictors measures such as true positive rate (TPR) or false negative rate (FNR) are used. In the literature [7], [5] a generalization to both continuous and categorical predictors can be found.

True Positive Rate is the proportion of true differences which are correctly identified, meaning ratio of the number of true elementary hypotheses which were found by the selector to the number of all true elementary hypotheses.

False Negative Rate is the proportion of false differences which are correctly identified, meaning ratio of the number of false elementary hypotheses which were rejected by the selector to the number of all false elementary hypotheses.

However, measures defined in this way diminish the influence of the continuous variables and factors with a small number of levels. As an example, consider a model with 5 continuous predictors and one factor with 5 levels. Then the number of parameters for the continuous predictors is 5 and the number of possible elementary hypotheses equals 5. The number of parameters for the categorical variable is also 5, whereas the number of possible elementary hypotheses is $\binom{5}{2} = 10$.

Therefore, we introduce a different generalization of traditional performance measures which treats the set of considered hypotheses as a linear subspace of the parameter space. The new measures are functions of sizes of the true and selected models. We consider two models: true model t and selected model s.

Definition 5. *Let us denote $t \cap s = \mathcal{L}(t) \cap \mathcal{L}(s)$.* Sensitivity *coefficient is defined as:*

$$Sen = \frac{TP}{FN + TP} = \frac{|t \cap s|}{|t|}.$$

Specificity *coefficient is defined as:*

$$Spe = \frac{TN}{FP + TN} = \frac{p - (|t| + |s| - |t \cap s|)}{p - |t|}.$$

However, in the article the attention is focused on values: $1 - Sen$ and $1 - Spe$, which correspond to the errors made by selector.

5.2 Measure of quality of selection

In all simulations described in the article we used BIC for model selection. One can ask how much better than the other models the selected model is. An answer to this question can be Bayes factors. Assuming uniform prior distribution on the set of models and denoting data by $\mathcal{D}$, the Bayes factor for the model m with respect to the best model $\widehat{m}$ (with minimum BIC) is expressed as

$$BF_m = \frac{\mathbb{P}(\mathcal{D}|m)}{\mathbb{P}(\mathcal{D}|\widehat{m})}.$$

Approximate Bayes factors [11]

$$\widetilde{BF}_m = \exp(-\frac{1}{2}(\mathrm{BIC}_m - \mathrm{BIC}_{\widehat{m}}))$$

are estimators of the quality of selection. Figure 3 illustrates an example of use of approximate Bayes factors for this purpose.

6 Simulation study

In order to compare DMR algorithm with other methods of model selection simulation studies were performed. All the simulations were conducted using functions implemented in `R` package called `DMR`, which is available at the `CRAN` webpage:

`http://cran.r-project.org/web/packages/DMR/index.html`

The main function of the package is `DMR`, which enables choosing a method of hierarchical clustering and a value of GIC penalty used by the algorithm. Moreover, other functions for extensions of DMR method such as `stepDMR`, which is based on recalculation of t-statistics in each step, and function `DMR4glm` for generalized linear models can be found. Functions `roc` and `plot_bf` can be used for obtaining measures of performance and quality of selection described in Section 5.

In section 6.1 results regarding an experiment described by Bondell and Reich, 2009, [5] are presented. The data generated for the experiment consists of three factors and no continuous variables. As a continuation, simulations based on data containing one factor and eight correlated continuous predictors were carried out. The results can be found in Section 6.2. In both experiments the complete linkage method of clustering in DMR algorithm and BIC were used. Section 6.4 focuses on comparison of this clustering method with others like single linkage and Ward's minimum variance methods. The last Section 6.3 refers to a real data example where fuel consumption of cars was modeled. The data `Cars93` comes from `R` package `MASS`.

6.1 Experiment 1

The experimental model consists of three factors having eight, four and three levels, respectively. The response y is generated from the model, which can be formulated in two equivalent ways. The first one uses the notation from the model formulation (1), the second one is an illustrative version, where we can see exactly the partitions of factors.

$$\begin{aligned}
y =&\beta_{00}\mathbf{1}_n + X_1\beta_1 + X_2\beta_2 + X_3\beta_3 + \varepsilon \\
=&2\cdot\mathbf{1}_n + X_1(0,-3,-3,-3,-3,-2,-2)^T + X_2(0,0,0)^T + X_3(0,0)^T + \varepsilon \\
=&V_1\alpha_1 + V_2\alpha_2 + V_3\alpha_3 + \varepsilon \\
=&V_1(2,2,-1,-1,-1,-1,0,0)^T + V_2(0,0,0,0)^T + V_3(0,0,0)^T + \varepsilon,
\end{aligned}$$

where X_i equals V_i with first column removed for $i = 1,2,3$ and $\varepsilon \sim \mathcal{N}(0,\mathbb{I}_n)$. A balanced design was used with k observations for each combination of factor levels, which gives $n = 96\cdot k$, $k = 1,2,4$.

The data was generated 1000 times. In the simulation study we compared four algorithms: DMR, stepDMR, CAS-ANOVA (`R`-code for CAS-ANOVA can be found at `http://www4.stat.ncsu.edu/ bondell/Software/CasANOVA/CasANOVA.R`) and stepBIC, which is a stepwise backward procedure implemented in `stepAIC` function from `R` package `MASS`.

The results are summarized in Table 1. True model (TM) represents the percentage of time the procedure chose the entirely correct model. Correct factors (CF) represents the percentage of time the non-significant factors were eliminated and the true factor was kept. TPR represents the average percentage of true differences found, whereas FNR represents the average percentage of false differences which were correctly identified. 1-Sen and 1-Spe are defined according to Definition 5, MSEP stands for mean squared error of prediction for new data and MD is mean dimension of the selected model, both with standard deviations.

The results of Experiment 1 indicate that DMR algorithms performed at least twice better than CAS-ANOVA in terms of choosing the true model. Our procedures chose approximately smaller models with dimension closer to the dimension of the underlying true model, whose number of parameters equals three. There are no significant differences between mean squared errors of prediction for all considered algorithms. The main conclusion, that DMR procedures choose models which are smaller and closer to the proper one, is supported by the obtained values of 1 - sensitivity and 1 - specificity, which indicates smaller errors made by our methods.

Table 1: Results of the simulation study, Experiment 1.

n	Algorithm	TM	CF	TPR	FNR	1-Sen	1-Spe	MSEP (±sd)	MD (±sd)
96	DMR	44	74	96	88	.1	.08	1.08±.17	3.5±.7
	stepDMR	44	74	96	88	.1	.08	1.08±.17	3.5±.7
	CAS-ANOVA	16	82	97	77	.06	.2	1.09±.17	4.8±1.7
	stepBIC	0	97	100	52	0	.51	1.08±.16	8.1±.4
192	DMR	67	83	99	93	.03	.04	1.03±.11	3.3±.6
	stepDMR	67	83	99	93	.03	.04	1.03±.11	3.3±.6
	CAS-ANOVA	32	92	100	86	.01	.13	1.05±.11	4.3±1.3
	stepBIC	0	99	100	53	0	.5	1.04±.11	8 ±.2
384	DMR	77	88	100	96	0	.03	1.02±.07	3.3±.5
	stepDMR	76	88	100	96	0	.03	1.02±.07	3.2±.5
	CAS-ANOVA	49	97	100	91	0	.08	1.03±.08	3.8±1
	stepBIC	0	99	100	53	0	.5	1.03±.07	8 ±.2

In Figure 3 an exemplary run of DMR algorithm is illustrated. Each row of the figure corresponds to a model on the nested path of models searched through by the algorithm. The left panel shows consecutive partitions of factors on the path of the algorithm: in the first row there is the full model, the last row shows the model containing only intercept. The true model is

$$t = (P_1 = \{\{1,2\},\{3,4,5,6\},\{7,8\}\}, P_2 = \{1,2,3,4\}, P_3 = \{1,2,3\}).$$

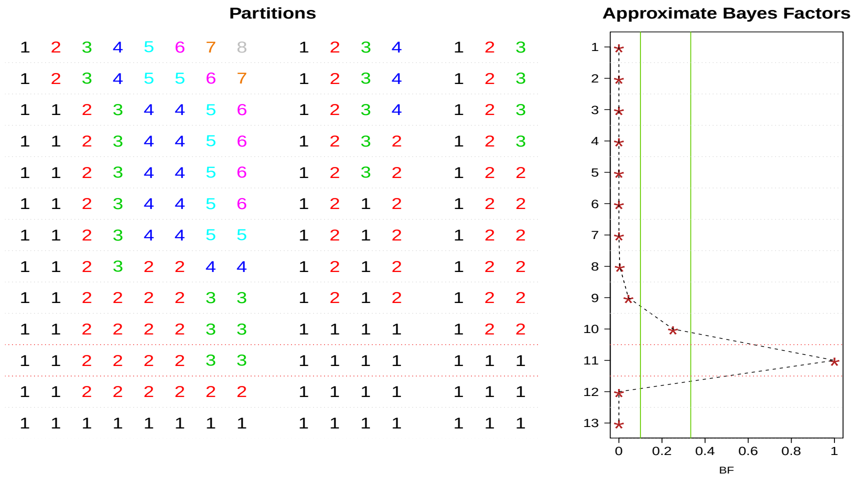
Bold dotted horizontal lines represent the final cut (a model with minimal BIC) of DMR algorithm. One can see that the best partition of the first factor consists of three groups, exactly the same as in the true effect vector, second and third factors are removed from the model.

The right panel of the figure shows approximate Bayes factors for models on the path. Two vertical lines represent the values of $\frac{1}{3}$ and $\frac{1}{10}$, which correspond to the Jeffreys scale [11] for interpretation of Bayes factors. We can see that there is at least substantial evidence to use the chosen model (with minimal BIC).

In Table 2 the results of computation times for several algorithms are summarized. All values are divided by the computation time of `lm.fit` function, which fits the linear model with the use of QR decomposition of the model matrix. The results for CAS-ANOVA are given only for one value of λ. By default, the searched lambda grid is of length 50. Hence, DMR is several hundred times faster than CAS-ANOVA.

6.2 Experiment 2

In the second simulation study a model containing not only categorical predictors, but also continuous variables is considered. The response y is generated from the



1	2	3	4	5	6	7	8	1	2	3	4	1	2	3
1	2	3	4	5	5	6	7	1	2	3	4	1	2	3
1	1	2	3	4	4	5	6	1	2	3	4	1	2	3
1	1	2	3	4	4	5	6	1	2	3	2	1	2	3
1	1	2	3	4	4	5	6	1	2	3	2	1	2	2
1	1	2	3	4	4	5	6	1	2	1	2	1	2	2
1	1	2	3	4	4	5	5	1	2	1	2	1	2	2
1	1	2	3	2	2	4	4	1	2	1	2	1	2	2
1	1	2	2	2	2	3	3	1	2	1	2	1	2	2
1	1	2	2	2	2	3	3	1	1	1	1	1	2	2
1	1	2	2	2	2	3	3	1	1	1	1	1	1	1
1	1	2	2	2	2	2	2	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Figure 3: An example run of DMR algorithm with Bayes factors.

Table 2: Computation times divided by the computation time of `lm.fit`.

k	n	DMR	CAS-ANOVA
1	96	35	122
3	288	15	45
21	2016	4	24

model:

$$\begin{aligned} y =& V_0\alpha_0 + V_1\alpha_1 + \varepsilon \\ =& V_0(1,0,1,0,1,0,1,0)^T + V_1(0,0,-2,-2,-2,-2,4,4)^T + \varepsilon, \end{aligned}$$

where V_0 was generated from the multivariate normal distributions with autoregressive correlation structure with $\rho = 0.8$. The first $2 \cdot 16 \cdot k$ observations come from a distribution with mean vector $(1,1,0,0,0,0,0,0)^T$, then $4 \cdot 16 \cdot k$ observations with mean vector $(0,0,1,1,1,1,0,0)^T$ and the last $2 \cdot 16 \cdot k$ observations with mean vector $(0,0,0,0,0,0,1,1)^T$, according to the underlying true partition of the factor, $k = 1,2,4$, hence $n = 128 \cdot k$. V_1 is matrix of dummy variables decoding levels of the factor and $\varepsilon \sim \mathcal{N}(0, \mathbb{I}_n)$.

Table 3 shows the results of simulation study. The data was generated 1000 times. As in Experiment 1 we compared four algorithms: DMR, stepDMR, CAS-ANOVA and stepBIC. Despite the additional continuous correlated variables the obtained results show a considerable advantage of DMR algorithms over other methods.

Note that in both Experiment 1 and Experiment 2 DMR algorithm having lower time complexity performed comparatively to stepDMR.

Table 3: Results of the simulation study, Experiment 2.

n	Algorithm	TM	1-Sen	1-Spe	MSEP (±sd)	MD (±sd)
128	DMR	69	0	.05	1.08±.14	7.4±.7
	stepDMR	68	0	.04	1.08±.14	7.4±.6
	CAS-ANOVA	15	.09	.31	1.11±.15	9.2±1.7
	stepBIC	0	0	.58	1.11±.15	12.2±.4
256	DMR	82	0	.02	1.03±.09	7.2±.5
	stepDMR	81	0	.02	1.03±.09	7.2±.5
	CAS-ANOVA	27	.1	.25	1.05±.1	8.5±1.4
	stepBIC	0	0	.57	1.04±.1	12.1±.3
512	DMR	86	0	.02	1.02±.06	7.1±.4
	stepDMR	86	0	.02	1.02±.06	7.1±.4
	CAS-ANOVA	43	.12	.2	1.03±.06	7.9±1.1
	stepBIC	0	0	.56	1.03±.06	12 ±.3

6.3 Real data example, `Cars93`

The data `Cars93` used in this example comes from `R` library `MASS`. 81 observations of 7 categorical and 15 continuous predictors are given. The logarithm of fuel consumption is under investigation. The factors are: presence of airbags, number of cylinders, drive train, availability of manual transmission version, origin, number of passengers and type. The continuous variables are: fuel tank capacity, length, logarithm of engine size, maximum horsepower, logarithm of engine revolutions per mile, luggage capacity, price, rear seat room, revs per minute at maximum horsepower, U-turn space, weight, wheelbase and width. These give 16 and 13 parameters respectively.

Model selection was performed using four methods: DMR, stepDMR, CAS-ANOVA and stepBIC. Characteristics of the chosen models are shown in Table 4 with results for the full model added for comparison. Figure 4 illustrates partitions of factors in the model selected by DMR procedure. From the set of continuous variables weight, wheelbase and logarithm of engine size were chosen.

Table 4: Cars93 data analysis results for different selection methods. Approximate Bayes factors calculated with respect to model chosen by stepDMR.

Selection	Number of	R^2	BIC	Bayes
method	parameters			factor
Full model	31	.92	-83.7	$2.7 \cdot 10^{-16}$
DMR	11	.9	-152.7	0.27
stepDMR	11	.9	-155.3	1
CAS-ANOVA	4	.8	-130.5	$4.1 \cdot 10^{-6}$
stepBIC	14	.89	-133.8	$2.1 \cdot 10^{-5}$

We can conclude that DMR procedures chose much better models than other compared methods in terms of BIC. Approximate Bayes factors for the full model and models chosen by stepBIC and CAS-ANOVA indicate the decisive evidence in favor of the model chosen by stepDMR according to Jeffrey's scale for interpretation of Bayes factors [11].

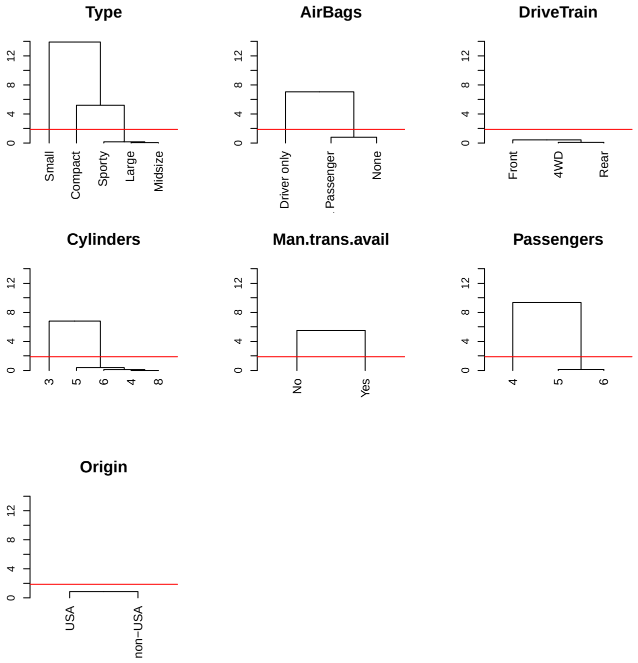


Figure 4: Partitions of factors for the model selected by DMR procedure for Cars93 data.

6.4 Clustering methods

DMR algorithm uses hierarchical clustering for generating a path of nested models. There is a wide spectrum of hierarchical clustering methods available in the statistical software. In order to compare some of them a simulation study was conducted. The results are summarized in Table 5. One can see that method complete gives the most stable results, therefore we decided to use it in simulation studies described in the article.

Table 5: Comparison of clustering methods for DMR algorithm for data as in Experiment 1.

n	Method	TM	1-Sen	1-Spe	MSEP (±sd)	MD (±sd)
96	complete	44	.11	.08	$1.09\pm.17$	$3.5\pm.7$
	single	36	.12	.09	$1.10\pm.18$	$3.6\pm.9$
	Ward's	44	.11	.08	$1.09\pm.17$	$3.5\pm.7$
192	complete	67	.03	.04	$1.03\pm.11$	$3.3\pm.5$
	single	65	.03	.04	$1.03\pm.11$	$3.4\pm.6$
	Ward's	67	.03	.04	$1.03\pm.11$	$3.3\pm.5$
384	complete	78	0	.02	$1.01\pm.07$	$3.2\pm.5$
	single	80	0	.02	$1.01\pm.07$	$3.2\pm.5$
	Ward's	78	0	.02	$1.01\pm.07$	$3.2\pm.5$

7 Conclusions

In this article novel methods of linear model selection combining deleting continuous variables with merging levels of factors were proposed. Both of them are based on ordering elementary hypotheses using squared t-statistics and choosing the best model according to GIC in the nested family of models.

We showed by simulations that DMR algorithms work well for small data sets. In comparison to other methods they gave much higher rates of choosing the true model. The time complexities of the algorithms are $O(np^2)$ and $O(np^3)$ for DMR and stepDMR respectively. In the simulations the algorithms worked several hundred times faster than for example CAS-ANOVA algorithm. For large data sets some asymptotic results were obtained. We proved that even under assumption that the number of predictors grows with the number of observations, $p_n \to \infty$, DMR algorithm is consistent.

Furthermore, a generalization of traditional measures of performance was introduced. These measures do not diminish the influence of factors with a small number of levels and continuous variables.

As a future work we plan to generalize the methods on several classes of models such as linear models with $p_n > n$, Generalized Linear Models and Cox models for survival data.

Acknowledgments

Study was supported by research fellowship within "Information technologies: research and their interdisciplinary applications" agreement number POKL.04.01.01-00-051/10-00.

Appendix

A Calculation of RSS for model with linear constraints

Let us consider a linear model:

$$\begin{cases} y = X\beta + \varepsilon \\ A_0\beta = c, \end{cases} \tag{6}$$

where A_0 is a $(p-q) \times p$ given matrix and c is a given vector of length $(p-q)$, which define a linear subspace of parameter space $\mathbb{R}^p$ described by a set of linear hypotheses concerning the vector of parameters β. The objective is to calculate residual sum of squares $\|y - X\widehat{\beta}_c\|^2$, where $\widehat{\beta}_c$ is an estimator of the parameter vector β with given constraints.

The following QR decomposition is performed

$$X = Q_p R,$$

where Q_p is $n \times p$ orthogonal matrix and R is $p \times p$ upper triangular matrix. Let us denote $S = R^{-T}A_0^T$, then

$$\begin{cases} Q_p^T y = R\beta + Q_p^T\varepsilon \\ S^T R\beta = c. \end{cases}$$

and after substitution we get

$$\begin{cases} z = \gamma + \eta \\ U^T W_{p-q}^T \gamma = c, \end{cases} \tag{7}$$

where W_{p-q} and U are respectively $p\times(p-q)$ orthogonal matrix and $(p-q)\times(p-q)$ upper triangular matrix from the QR decomposition of matrix S. If we denote

$$W_{p-q}^T\gamma = U^{-T}c = e$$

and

$$W = [W_q, W_{p-q}],$$

where W_q is an orthogonal complement of the matrix W_{p-q}, then equation (7) becomes

$$\begin{cases} W_q^T z = W_q^T\gamma + W_q^T\eta \\ e = W_{p-q}^T\gamma. \end{cases}$$

Therefore an unbiased estimator of γ with constraints satisfies the following equation

$$\begin{bmatrix} W_q^T z \\ e \end{bmatrix} = W^T\widehat{\gamma}_c, \tag{8}$$

multiplying (8) by W, we obtain

$$W_q W_q^T z + W_{p-q}e = \widehat{\gamma}_c,$$

then

$$(I_p - W_{p-q}W_{p-q}^T)z + W_{p-q}e = \widehat{\gamma}_c = R\widehat{\beta}_c.$$

The residual sum of squares for the model with linear constraints (6) can be written as

$$\begin{aligned}\|y - X\widehat{\beta}_c\|^2 =& \|Q_p^T y - R\widehat{\beta}_c\|^2 = \|W_{p-q}W_{p-q}^T z - W_{p-q}e\|^2 \\ =&\|W_{p-q}^T z - e\|^2 = \sum_{i=1}^{p-q}(w_i^T z - e_i)^2,\end{aligned} \tag{9}$$

where w_i is the i-th column of matrix W_{p-q}. Hence, for each additional hypothesis the residual sum of squares can be easily calculated from equation (9).

B Proof of Theorem 1

Lemma 1 (Klotz [12] Section 14.3)**.** *Solving an optimization problem:*

$$\widehat{\beta}_0 = \arg\min_{\beta} ||y - X\beta||^2, \textit{ where } A_0\beta = 0 \textit{ yields}$$

$$\widehat{\beta}_0 = \widehat{\beta} - (X^TX)^{-1}A_0^T(A_0(X^TX)^{-1}A_0^T)^{-1}A_0\widehat{\beta},$$

where $\widehat{\beta} = (X^TX)^{-1}X^Ty$.

Let us denote the following matrices:

$$H = X(X^TX)^{-1}X^T,$$

$$H_0 = X(X^TX)^{-1}A_0^T(A_0(X^TX)^{-1}A_0^T)^{-1}A_0(X^TX)^{-1}X^T$$

which are matrices of orthogonal projections and matrix

$$H_1 = Z_1(Z_1^TZ_1)^{-1}Z_1^T = XA^1({A^1}^TX^TXA^1)^{-1}{A^1}^TX^T.$$

Note that $X\widehat{\beta}_0 = (H - H_0)y$.

Lemma 2. *H_1 is a matrix of an orthogonal projection and*

$$H_1 = H - H_0.$$

Proof. Note that

$$(X^TX)^{-1} = (A^TA^{-T}X^TXA^{-1}A)^{-1} = A^{-1}(A^{-T}X^TXA^{-1})^{-1}A^{-T},$$

hence

$$\begin{aligned}(A_0(X^TX)^{-1}A_0^T)^{-1} =& (A_0A^{-1}(A^{-T}X^TXA^{-1})^{-1}A^{-T}A_0^T)^{-1} \\ =& \left[\begin{bmatrix} 0 & \mathbb{I}\end{bmatrix}(Z^TZ)^{-1}\begin{bmatrix} 0 \\ \mathbb{I}\end{bmatrix}\right]^{-1} = (G^{00})^{-1},\end{aligned}$$

where

$$G = \begin{bmatrix} G_{11} & G_{10} \\ G_{01} & G_{00}\end{bmatrix} = \begin{bmatrix} Z_1^TZ_1 & Z_1^TZ_0 \\ Z_0^TZ_1 & Z_0^TZ_0\end{bmatrix} = Z^TZ \text{ and } G^{-1} = \begin{bmatrix} G^{11} & G^{10} \\ G^{01} & G^{00}\end{bmatrix}$$

and

$$\begin{aligned} H =&X(X^TX)^{-1}X^T \\ =&XA^{-1}(A^{-T}X^TXA^{-1})^{-1}A^{-T}X^T = Z(Z^TZ)^{-1}Z^T. \end{aligned}$$

We have also that

$$A_0(X^TX)^{-1}X^T = A_0A^{-1}(Z^TZ)^{-1}A^{-T}X^T = A_0A^{-1}(Z^TZ)^{-1}Z^T,$$

then

$$\begin{aligned} H_0 =&X(X^TX)^{-1}A_0^T(G^{00})^{-1}A_0(X^TX)^{-1}X^T \\ =&Z(Z^TZ)^{-1}\begin{bmatrix} 0 \\ \mathbb{I} \end{bmatrix}(G^{00})^{-1}\begin{bmatrix} 0 & \mathbb{I} \end{bmatrix}(Z^TZ)^{-1}Z^T \\ =&Z\begin{bmatrix} G^{10} \\ G^{00} \end{bmatrix}(G^{00})^{-1}\begin{bmatrix} G^{01} & G^{00} \end{bmatrix}Z^T = Z\begin{bmatrix} G^{10}(G^{00})^{-1}G^{01} & G^{10} \\ G^{01} & G^{00} \end{bmatrix}Z^T. \end{aligned}$$

Using above results we get

$$\begin{aligned} H - H_0 =&Z(Z^TZ)^{-1}Z^T - Z\begin{bmatrix} G^{10}(G^{00})^{-1}G^{01} & G^{10} \\ G^{01} & G^{00} \end{bmatrix}Z^T \\ =&\begin{bmatrix} Z_1 & Z_0 \end{bmatrix}\begin{bmatrix} G^{11} - G^{10}(G^{00})^{-1}G^{01} & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} Z_1^T \\ Z_0^T \end{bmatrix} \\ =&Z_1(G^{11} - G^{10}(G^{00})^{-1}G^{01})Z_1^T. \end{aligned}$$

From the definition of H_1

$$H_1 = Z_1(Z_1^TZ_1)^{-1}Z_1^T = Z_1(G_{11})^{-1}Z_1^T \tag{10}$$

and from Schur complement for matrix G^{-1} we have

$$(G_{11})^{-1} = G^{11} - G^{10}(G^{00})^{-1}G^{01}.$$

□

Hence, the predictions for constrained problem can be obtained through projecting the observations on the space spanned by columns of the model matrix for the equivalent unconstrained problem.

B.1 RSS lemmas

Lemmas concerning dependencies between residual sums of squares have similar construction to those described by Chen and Chen, 2008 [13].

It follows from Lemma 2 that for each feasible model m a hat matrix H_m can be obtained according to equation (10), so that residual sum of squares for model m is defined as

$$RSS_m = \|y - H_my\|^2$$

and can be decomposed into three parts

$$\begin{aligned} RSS_m = \|y - H_m y\|^2 =& (X_t\beta_t + \varepsilon)^T(\mathbb{I} - H_m)(X_t\beta_t + \varepsilon) \\ =& \beta_t^T X_t^T(\mathbb{I} - H_m)X_t\beta_t + 2\beta_t^T X_t^T(\mathbb{I} - H_m)\varepsilon + \varepsilon^T H_m\varepsilon. \end{aligned}$$

When $t \subseteq m$ we have $H_m X_t = X_t$ and

$$RSS_m = \varepsilon^T H_m \varepsilon.$$

Lemma 3. *Assuming $p_n \prec n$ and r_n is a sequence of real numbers so that $p_n \prec r_n$, then*

$$\log \frac{RSS_t}{RSS_f} <_P \frac{r_n}{n},$$

which signifies $\lim_{n\to\infty} \mathbb{P}\left(\log \frac{RSS_t}{RSS_f} < \frac{r_n}{n}\right) = 1.$

Proof. Note that

$$\frac{RSS_t}{RSS_f} = 1 + \frac{RSS_t - RSS_f}{RSS_f} = 1 + \frac{p_n}{n}W_n,$$

where

$$W_n = \frac{\varepsilon^T(H_f - H_t)\varepsilon}{\varepsilon^T(\mathbb{I} - H_f)\varepsilon} \cdot \frac{n}{p_n}.$$

From Lemma 2, $H_f - H_t$ is matrix of an orthogonal projection with trace $p_n - |t|$, therefore $\psi_1 = \varepsilon^T(H_f - H_t)\varepsilon \sim \sigma^2\chi^2_{p_n - |t|}$ and $\psi_2 = \varepsilon^T(\mathbb{I} - H_f)\varepsilon \sim \sigma^2\chi^2_{n-p_n}$, we get

$$\mathbb{E}\left(\frac{\psi_1}{p_n}\right) = \frac{\sigma^2(p_n - |t|)}{p_n}, \ \operatorname{Var}\left(\frac{\psi_1}{p_n}\right) = \frac{2\sigma^4(p_n - |t|)}{p_n^2}$$

and either if $p_n \xrightarrow{n\to\infty} \infty$ then $\operatorname{Var}\left(\frac{\psi_1}{p_n}\right) \xrightarrow{n\to\infty} 0$ and from Chebyshev's inequality $\frac{\psi_1}{p_n} \xrightarrow{n\to\infty} \sigma^2$ in probability or if p_n is bounded, then $\frac{\psi_1}{p_n}$ is bounded in probability. Analogously for ψ_2 we have

$$\mathbb{E}\left(\frac{\psi_2}{n}\right) = \frac{\sigma^2(n - p_n)}{n}, \ \operatorname{Var}\left(\frac{\psi_2}{n}\right) = \frac{2\sigma^4(n - p_n)}{n^2}$$

and since $p_n \prec n$ from Chebyshev's inequality $\frac{\psi_2}{n} \xrightarrow{n\to\infty} \sigma^2$ in probability.

Therefore $W_n = O_P(1)$ and $\frac{RSS_t}{RSS_f} = 1 + O_P\left(\frac{p_n}{n}\right)$. Hence

$$\log\left(\frac{RSS_t}{RSS_f}\right) = \log\left(1 + \frac{p_n}{n}W_n\right) \le \frac{p_n}{n}W_n = O_P\left(\frac{p_n}{n}\right) <_P \frac{r_n}{n}.$$

□

Lemma 4. *Assuming that $p_n \prec d_n$ we have for all $m \in \mathcal{M}_\mathcal{V}$ and all $\delta > 1$*

$$\log\left(\frac{RSS_m}{RSS_t}\right) \ge_P \log\left(1 + \frac{d_n}{\delta\sigma^2 \cdot n}\right),$$

where $d_n = \min_{m\in\mathcal{M}_\mathcal{V}} \beta_t^T X_t^T(\mathbb{I} - H_m)X_t\beta_t$.

Proof. Using the fact that

$$\frac{1}{n}RSS_t = \frac{\varepsilon^T(\mathbb{I} - H_t)\varepsilon}{n} = \sigma^2 + o_P(1)$$

and denoting

$$RSS_m - RSS_t = d_{nm} + Z_{nm} + W_{nt} - W_{nm},$$

where $d_{nm} = \beta_t^T X_t^T(\mathbb{I} - H_m)X_t\beta_t$, $Z_{nm} = 2\beta_t^T X_t^T(\mathbb{I} - H_m)\varepsilon$, $W_{nt} = \varepsilon^T H_t\varepsilon$ and $W_{nm} = \varepsilon^T H_m\varepsilon$.

Note that

$$d_{nm} \geq d_n,\ Z_{nm} \sim \mathcal{N}(0, 4\sigma^2 d_{nm}),\ W_{nt} \sim \sigma^2\chi^2_{|t|} \text{ and } W_{nm} \sim \sigma^2\chi^2_{p_n-1}.$$

Using assumptions, $\frac{Z_{nm}}{d_{nm}}$, $\frac{W_{nt}}{d_{nm}}$ and $\frac{W_{nm}}{d_{nm}}$ either are bounded in probability if p_n is bounded or if $p_n \xrightarrow{n\to\infty} \infty$ are $O_P(1)$ from Chebyshev's inequality.

Henceforth we have

$$RSS_m - RSS_t = d_{nm}\left(1 + \frac{Z_{nm}}{d_{nm}} + \frac{W_{nt}}{d_{nm}} - \frac{W_{nm}}{d_{nm}}\right) = d_{nm}\Big(1 + O_P(1)\Big).$$

As a result

$$\begin{aligned}\log\frac{RSS_m}{RSS_t} &= \log\left(1 + \frac{RSS_m - RSS_t}{n\frac{RSS_t}{n}}\right)\\ &= \log\left(1 + \frac{d_{nm}}{n\sigma^2}\Big(1 + O_P(1)\Big)\right) \geq_P \log\left(1 + \frac{d_n}{\delta\sigma^2 n}\right) \text{ for } \delta > 1.\end{aligned}$$

□

Lemma 5. *Assuming that $r_n \prec n$ for all $\delta > 1$ we have*

$$\max_{t\subseteq m\subseteq\mathcal{M}}\Big(\log RSS_m\Big) + \log\left(1 + \frac{d_n}{\delta\sigma^2 n}\right) \leq_P \min_{t\nsubseteq m\subseteq\mathcal{M}}\Big(\log RSS_m\Big).$$

Proof. Let us denote $a = \log\left(1 + \frac{d_n}{\delta\sigma^2 n}\right)$, then from Lemma 4 we get

$$\begin{aligned}\max_{t\subseteq m\subseteq\mathcal{M}}\Big(\log RSS_m\Big) + a &= \log RSS_t + a \leq_P \min_{m\in\mathcal{M}_\mathcal{V}}\log RSS_m - a + a\\ &\leq_P \min_{t\nsubseteq m\subseteq\mathcal{M}}\Big(\log RSS_m\Big).\end{aligned}$$

□

Corollary 1. *From Lemma 5 and properties of residual sum of squares, we have with probability tending to 1 the following order of models RSS:*

$$\begin{aligned}RSS_f &\leq \max_{|m|=|f|-1,\ t\subseteq m} RSS_m \leq \max_{|m|=|f|-2,\ t\subseteq m} RSS_m\\ &\leq \ldots \leq \max_{|m|=|t|+1,\ t\subseteq m} RSS_m \leq RSS_t \leq_P \min_{|m|=|f|-1,\ t\nsubseteq m} RSS_m\\ &\leq \min_{|m|=|f|-2,\ t\nsubseteq m} RSS_m \leq \ldots \leq \min_{|m|=2,\ t\nsubseteq m} RSS_m \leq \min_{|m|=1} RSS_m.\end{aligned} \tag{11}$$

Note that since $|t| < \infty$, there is a finite number of models bigger and having RSS not greater than the true model t.

Corollary 2. *Since models from $\mathcal{M}_\mathcal{V}$ and $\mathcal{M}_\mathcal{T}$ have the same number of parameters, we have*

$$GIC(t) \leq_P \max_{m \in \mathcal{M}_\mathcal{T}} GIC(m) \leq_P \min_{m \in \mathcal{M}_\mathcal{V}} GIC(m).$$

where the left inequality follows form Lemma 3 and right inequality from Lemma 5.

Corollary 3. *For every step of a backward stepwise elimination algorithm, if the true model t is on the path searched through, the inequality from Corollary 2 is preserved. Hence, GIC is a consistent model selection criterion.*

B.2 Ordering of squared t-statistics

Lemma 6. *Suppose $\mathcal{M}_{\mathcal{T}\mathcal{V}} = \mathcal{M}_\mathcal{T} \cup \mathcal{M}_\mathcal{V}$ is a set of all models of size $|f| - 1$. For each $m \in \mathcal{M}_{\mathcal{T}\mathcal{V}}$, which corresponds to one elementary hypothesis,*

$$T_m^2 = (n - |f|)\frac{RSS_m - RSS_f}{RSS_f},$$

where T_m is t-statistic for the full model with hypothesis $h: \ A_0\beta = 0$, where A_0 is $1 \times |f|$ matrix.

Proof. From Lemma 1

$$RSS_f - RSS_m = \widehat{\beta}^T A_0^T (A_0 (X^T X)^{-1} A_0^T)^{-1} A_0 \widehat{\beta},$$

hence

$$T_m^2 = \frac{(A_0\widehat{\beta})^2}{\widehat{\mathrm{Var}}(A_0\widehat{\beta})} = \frac{(A_0\widehat{\beta})^2}{A_0\widehat{\mathrm{Var}}(\widehat{\beta})A_0^T} = \frac{(A_0\widehat{\beta})^2}{\widehat{\sigma}^2 A_0 (X^T X)^{-1} A_0^T} = \frac{RSS_f - RSS_m}{\widehat{\sigma}^2},$$

where $\widehat{\sigma}^2 = \frac{RSS_f}{n-|f|}$. □

Corollary 4. *It follows from Lemma 6 that the ordering of models $m \in \mathcal{M}_{\mathcal{T}\mathcal{V}}$ with respect to squared t-statistics is equivalent to ordering them with respect to the values of residual sum of squares for these models.*

Corollary 5. *It follows from Corollary 1 and Lemma 6, that for sufficiently large n we have*

$$\max_{m \in \mathcal{M}_\mathcal{T}} T_m^2 <_P c <_P \min_{m \in \mathcal{M}_\mathcal{V}} T_m^2,$$

where c is a positive constant.

In order to prove that hierarchical clustering implies the proper order of accepting elementary hypotheses, in the sense that the true hypotheses preface false ones, let us introduce some notations.

Let $\mathcal{T}$ stand for a set of all true elementary hypotheses and $\mathcal{V}$ for a set of all false elementary hypotheses. Suppose that $U_1^{(s)}, U_2^{(s)}, \ldots U_{k_s}^{(s)}$ are clusters of levels of a factor given by the s-th step of hierarchical clustering. $\beta_1^{(s)}, \beta_2^{(s)}, \ldots \beta_{k_s}^{(s)}$ correspond to the common value of the parameters for the factor levels for the clusters. We assume that each $\beta_i^{(s)}$ is defined as the parameter for the level with the smallest index in the cluster.

Lemma 7. *In each step of hierarchical clustering in DMR algorithm the recalculation of dissimilarity matrix using linkage criterion, which is a convex combination of minimum and maximum of the pairwise distances between clusters, preserves the inequality:*

$$\max_{\{\beta_{i_1}^{(s)}=\beta_{i_2}^{(s)}\}\in\mathcal{T}} d\left(U_{i_1}^{(s)}, U_{i_2}^{(s)}\right) <_P c <_P \min_{\{\beta_{i_1}^{(s)}=\beta_{i_2}^{(s)}\}\in\mathcal{V}} d\left(U_{i_1}^{(s)}, U_{i_2}^{(s)}\right),$$

where d is the dissimilarity measure and $c > 0$.

Proof. By induction: for $s = 1$ the inequality is preserved from Corollary 5 and the hypothesis with the smallest value of squared t-statistic is accepted. Let us assume that the inequality is also preserved for step s.

If $s + 1 > |\mathcal{T}|$ all true hypotheses are already accepted, so the inequality holds trivially for all following steps.

If $s + 1 \leq |\mathcal{T}|$ the algorithm of clustering chooses the hypothesis with minimal value of dissimilarity measure to accept. Suppose that this hypothesis has a form

$$\beta_{i_1}^{(s)} = \beta_{i_2}^{(s)}.$$

From the previous step s we know that this hypothesis is true. The dissimilarity measures between the new merged cluster and every other cluster $U_{i_3}^{(s)}$ have to be recalculated. We have two cases:

1. If $\beta_{i_3}^{(s)} = \beta_{i_1}^{(s)}$ is true then from transitivity $\beta_{i_3}^{(s)} = \beta_{i_2}^{(s)}$ is also true, hence

$$d\left(U_{i_3}^{(s)}, U_{i_1}^{(s)}\right) = d_{i_3 i_1}^{(s)} <_P c,$$

$$d\left(U_{i_3}^{(s)}, U_{i_2}^{(s)}\right) = d_{i_3 i_2}^{(s)} <_P c$$

so the convex combination of these two values is smaller than c:

$$\begin{aligned} & d\left(\{U_{i_1}^{(s)}, U_{i_2}^{(s)}\} = U_{j_1}^{(s+1)}, U_{i_3}^{(s)} = U_{j_2}^{(s+1)}\right) \\ & = \alpha \cdot \min\left(d_{i_3 i_1}^{(s)}, d_{i_3 i_2}^{(s)}\right) + (1-\alpha) \cdot \max\left(d_{i_3 i_1}^{(s)}, d_{i_3 i_2}^{(s)}\right) <_P c \end{aligned}$$

for each $\alpha \in [0, 1]$.

2. Analogously if $\beta_{i_3}^{(s)} = \beta_{i_1}^{(s)}$ is false the convex combination of these two values is greater than c:

$$\begin{aligned} & d\left(\{U_{i_1}^{(s)}, U_{i_2}^{(s)}\} = U_{j_1}^{(s+1)}, U_{i_3}^{(s)} = U_{j_2}^{(s+1)}\right) \\ & = \alpha \cdot \min\left(d_{i_3 i_1}^{(s)}, d_{i_3 i_2}^{(s)}\right) + (1-\alpha) \cdot \max\left(d_{i_3 i_1}^{(s)}, d_{i_3 i_2}^{(s)}\right) >_P c \end{aligned}$$

for each $\alpha \in [0,1]$. Note that linkage criteria: single, complete and average are a convex combination of minimum and maximum of the pairwise distances between clusters.

□

Proof of Theorem 1. It follows from Corollary 5 and Lemma 7 that for sufficiently large n on the path of models generated by DMR algorithm models with only true hypotheses accepted preface models with at least one false hypothesis accepted. Hence the true model t is on the path searched through. Therefore, from Corollary 3 we have that DMR algorithm is a consistent model selection method.

References

[1] Tukey, J.W.: Comparing individual means in the analysis of variance. Biometrics (1949) 99–114

[2] Scott, A., Knott, M.: A cluster analysis method for grouping means in the analysis of variance. Biometrics (1974) 507–512

[3] Calinski, T., Corsten, L.: Clustering means in anova by simultaneous testing. Biometrics (1985) 39–48

[4] Corsten, L., Denis, J.: Structuring interaction in two-way tables by clustering. Biometrics (1990) 207–215

[5] Bondell, H.D., Reich, B.J.: Simultaneous factor selection and collapsing levels in anova. Biometrics **65**(1) (2009) 169–177

[6] Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological) (1996) 267–288

[7] Gertheiss, J., Tutz, G.: Sparse modeling of categorial explanatory variables. The Annals of Applied Statistics **4**(4) (2010) 2150–2180

[8] Efron, B., Hastie, T., Johnstone, I., Tibshirani, R.: Least angle regression. The Annals of Statistics **32**(2) (2004) 407–499

[9] Ciampi, A., Lechevallier, Y., Limas, M.C., Marcos, A.G.: Hierarchical clustering of subpopulations with a dissimilarity based on the likelihood ratio statistic: application to clustering massive data sets. Pattern Analysis & Applications **11**(2) (2008) 199–220

[10] Zheng, X., Loh, W.Y.: Consistent variable selection in linear models. Journal of the American Statistical Association **90**(429) (1995) 151–156

[11] Kass, R.E., Raftery, A.E.: Bayes factors. Journal of the American Statistical Association **90**(430) (1995) 773–795

[12] Klotz, J.H.: A computational approach to statistics. Department of Statistics, University of Wisconsin, Madison (2006)

[13] Chen, J., Chen, Z.: Extended bayesian information criteria for model selection with large model spaces. Biometrika **95**(3) (2008) 759–771