Electronic Journal of Statistics
Vol. 9 (2015) 1749–1778
ISSN: 1935-7524
DOI: 10.1214/15-EJS1050

Delete or merge regressors for linear model selection

Aleksandra Maj-Kańska[*]

Institute of Computer Science
Polish Academy of Sciences
Jana Kazimierza 5
01-248 Warsaw
Poland
e-mail: a.maj@phd.ipipan.waw.pl

Piotr Pokarowski[†]

Faculty of Mathematics, Informatics and Mechanics
University of Warsaw
Banacha 2
02-097 Warsaw
Poland
e-mail: pokar@mimuw.edu.pl

and

Agnieszka Prochenka[*]

Institute of Computer Science
Polish Academy of Sciences
Jana Kazimierza 5
01-248 Warsaw
Poland
e-mail: a.prochenka@phd.ipipan.waw.pl

Abstract: We consider a problem of linear model selection in the presence of both continuous and categorical predictors. Feasible models consist of subsets of numerical variables and partitions of levels of factors. A new algorithm called delete or merge regressors (DMR) is presented which is a stepwise backward procedure involving ranking the predictors according to squared t-statistics and choosing the final model minimizing BIC. We prove consistency of DMR when the number of predictors tends to infinity with the sample size and describe a simulation study using a pertaining R package. The results indicate significant advantage in time complexity and selection accuracy of our algorithm over Lasso-based methods described in the literature. Moreover, a version of DMR for generalized linear models is proposed.

MSC 2010 subject classifications: Primary 62F07; secondary 62J07.
Keywords and phrases: ANOVA, consistency, BIC, merging levels, t-statistic, variable selection.

Received May 2015.

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

[†] Study was supported by Polish National Science Center grant 2011/01/B/NZ2/00864.

Contents

1. Introduction

Model selection is usually understood as selection of continuous 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 of examining the relationship between a continuous response and categorical variables is analysis of variance (ANOVA). After detecting the overall importance of a factor, pairwise comparisons of group means are used to test significance of differences between its levels. Typically post-hoc analysis such as Tukey's honestly significant difference (HSD) test or multiple comparison adjustments (Bonferroni, Scheffe) are used. A drawback of pairwise comparisons is non-transitivity of conclusions.

For example, let us consider data `barley` from `R` library `lattice` discussed already in Bondell and Reich (2009). Total yield of barley for 5 varieties at 6 sites in each of two years is modeled. The dependence between the response and the varieties variable with the use of Tukey's HSD analysis (Figure 1) gives inconclusive answers: $\beta_P = \beta_M$, $\beta_P = \beta_T$, but $\beta_T \neq \beta_M$.

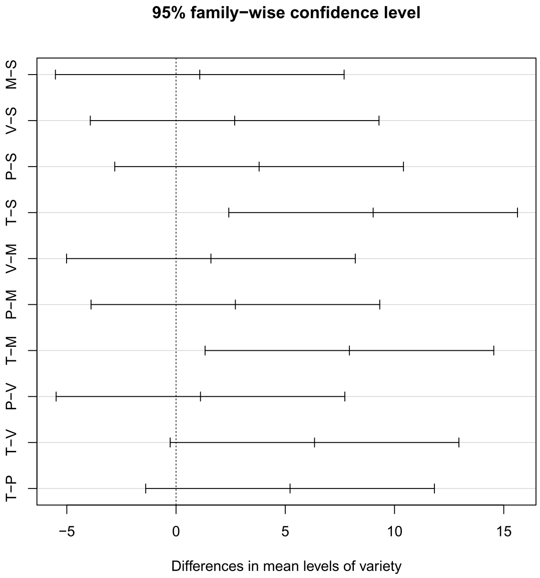

Fig 1. *Results of Tukey's HSD.*

In this work we introduce a novel procedure called delete or merge regressors (DMR), which enables an efficient search among partitions of factor levels, for which the issue of non-transitivity does not occur. If we apply DMR to the `barley` data, we get the following partition of varieties: $\{\{S, M, V, P\}, \{T\}\}$. Detailed description of the data set and the characteristics of the chosen model can be found in Section 5.5.

The idea of partitioning a set of levels of a factor into non-overlapping groups has already been discussed in the literature. In the article Tukey (1949) a stepwise backward procedure based on the studentized range which gives grouping of means for samples from normal distributions was proposed. Other methods of clustering of sample means were described in Scott and Knott (1974), where the set of means is partitioned from coarsest to finest, and in Caliński and Corsten (1985) whose algorithm adapts hierarchical clustering to the problem. In more recent articles Porreca and Ferrari-Trecate (2010) and Ciampi et al. (2008) efficient algorithms for datasets partitioning using generalized likelihood ratio test can be found. However, all the mentioned methods assume an arbitrary choice of significance level for the underlying tests. In our procedure we avoid the problem by selecting the final partition according to the minimal value of information criterion.

Information criterion as an objective function for partition selection is used in the procedures described in Dayton (2003). Dayton's SAS procedure, called paired comparisons information criteria (PCIC), computes AIC and BIC values for all ordered subsets of independent means for both homogeneous and heterogeneous models. In contrast to DMR these methods do not allow for simultaneous factor partitioning and selection of continuous variables.

A method introduced in Bondell and Reich (2009) called collapsing and shrinkage ANOVA (CAS-ANOVA) solves the same problem as DMR with the use of the least absolute shrinkage and selection operator (Lasso; Tibshirani (1996)), where the L_1 penalty is imposed on differences between parameters corresponding to levels of each factor. This algorithm can be interpreted as a generalization of fused Lasso (Tibshirani et al. (2004)) to data with categorical variables. In Gertheiss and Tutz (2010) one can find a modification of CAS-ANOVA, which is more computationally efficient because of using the least angle regression algorithm (LARS; Efron et al. (2004)). Another algorithm, based on regularized model selection with categorical predictors and effect modifiers (Oelker, Gertheiss and Tutz (2014)) is implemented in `R` package `gvcm.cat`. It generalizes the Lasso approach to simultaneous factor partitioning and selection of continuous variables to generalized linear models. The algorithm is based on local quadratic approximation and iterated reweighted least squares.

We propose a backward selection procedure called delete or merge regressors (DMR), which combines deleting continuous variables with merging levels of factors. The method employs a 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. In each step the choice of constraint is based on the order of squared t-statistics. As a result a nested family of linear models is obtained and the final decision is made by minimization of Bayesian information criterion (BIC). The method adapts agglomerative clustering, where squared t-statistics define the dissimilarity measure. This procedure generalizes concepts introduced in Zheng and Loh (1995) and Ciampi et al. (2008).

In the article we show that the DMR algorithm is a consistent model selection method under rather weak assumptions when p tends to infinity with n. Furthermore, thanks to using a recursive formula for RSS in a nested family of linear models, the time complexity of the DMR algorithm is just $O(np^2)$. This makes the algorithm much faster than the competitive Lasso-based methods. In the article we describe a simulation study and discuss a pertaining `R` package. The simulations show that DMR in comparison to adaptive Lasso methods described in the literature gives better results in terms of accuracy without the troublesome choice of the λ grid.

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

2. Feasible models

In this section we first introduce some definitions regarding the form of the data and models considered. In particular, we define the set of feasible models, which are linear spaces of parameters with linear constraints and we show how by change of variables the constrained problem can be replaced by an unconstrained one. Later we indicate that properties of OLS (ordinary least squares) estimators transfer to feasible models.

2.1. Definitions

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

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon} = \mathbf{1}\beta_{00}^* + \mathbf{X}_0\boldsymbol{\beta}_0^* + \mathbf{X}_1\boldsymbol{\beta}_1^* + \ldots + \mathbf{X}_l\boldsymbol{\beta}_l^* + \boldsymbol{\varepsilon}, \tag{1}$$

where:

1. $\boldsymbol{\varepsilon}$ is a vector of iid zero-mean gaussian errors, $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbb{I})$.
2. $\mathbf{X} = [\mathbf{1}, \mathbf{X}_0, \mathbf{X}_1, \ldots, \mathbf{X}_l]$ is a model matrix organized as follows: $\mathbf{X}_0$ is a matrix corresponding to continuous regressors and $\mathbf{X}_1, \ldots, \mathbf{X}_l$ are zero-one matrices encoding corresponding factors with the first level set as the reference.
3. $\boldsymbol{\beta}^* = [\beta_{00}^*, \boldsymbol{\beta}_0^{*T}, \boldsymbol{\beta}_1^{*T}, \ldots, \boldsymbol{\beta}_l^{*T}]^T \in \mathbb{R}^p$ is a parameter vector organized as follows: β_{00}^* is the intercept, $\boldsymbol{\beta}_0^* = [\beta_{10}^*, \ldots, \beta_{p_0 0}^*]^T$ is a vector of coefficients for continuous variables and $\boldsymbol{\beta}_k^* = [\beta_{2k}^*, \ldots, \beta_{p_k k}^*]^T$ is a vector of parameters corresponding to the k-th factor, $k = 1, \ldots, l$, hence the length of the parameter vector is $p = 1 + p_0 + (p_1 - 1) + \ldots + (p_l - 1)$.

Denote sets of indexes: $N = \{0, 1, \ldots, l\}$, $N_0 = \{0, 1, \ldots, p_0\}$ and $N_k = \{2, 3, \ldots, p_k\}$ for $k \in N \setminus \{0\}$. Let us define an elementary constraint for linear model (1) as a linear constraint of one of two types:

$$\mathcal{H}_{jk}: \ \beta_{jk}^* = 0 \text{ where } j \in N_k \setminus \{0\}, \ k \in N, \tag{2}$$

$$\mathcal{H}_{ijk}: \ \beta_{ik}^* = \beta_{jk}^* \text{ where } i, j \in N_k, \ i \neq j, \ k \in N \setminus \{0\}. \tag{3}$$

A feasible model can be defined as a sequence $M = (P_0, P_1, ..., P_l)$, where P_0 denotes a subset of indexes of continuous variables and P_k is a particular partition of levels of the k-th factor. Such a model can be encoded by a set of elementary constraints. A set of all feasible models is denoted by $\mathcal{M}$. Let us denote a model $F \in \mathcal{M}$ without constraints of types (2) or (3) as the full model.

Example 1. For illustration, let us consider a model with one factor and one continuous variable:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon} = \mathbf{1} \cdot 1 + \mathbf{X}_0 \cdot 2 + \mathbf{X}_1 \cdot \begin{bmatrix} -2 \\ -2 \\ 0 \end{bmatrix} + \boldsymbol{\varepsilon}$$

$$= \begin{bmatrix} 1\\1\\1\\1\\1\\1\\1\\1 \end{bmatrix} \cdot 1 + \begin{bmatrix} -0.96\\-0.29\\0.26\\-1.15\\0.2\\0.03\\0.09\\1.12 \end{bmatrix} \cdot 2 + \begin{bmatrix} 0&0&0\\0&0&0\\1&0&0\\1&0&0\\0&1&0\\0&1&0\\0&0&1\\0&0&1 \end{bmatrix} \begin{bmatrix} -2\\-2\\0 \end{bmatrix} + \begin{bmatrix} -1.22\\1.27\\-0.74\\-1.13\\-0.72\\0.25\\0.15\\-0.31 \end{bmatrix}, \quad (4)$$

where $\mathbf{X}_0$ and $\boldsymbol{\varepsilon}$ are vectors of length 8 generated independently from standard normal distribution, $\mathcal{N}(\mathbf{0}, \mathbb{I})$. Then $\beta^* = [1, 2, -2, -2, 0]^T$. The full model $F = (P_0 = \{1\}, P_1 = \{\{1\}, \{2\}, \{3\}, \{4\}\})$ with $p_0 = 1, p_1 = 4, p = 5$. The model corresponding to β^* is $(P_0 = \{1\}, P_1 = \{\{1, 4\}, \{2, 3\}\})$ and is the same as F with two elementary constraints: $\beta^*_{41} = 0$ and $\beta^*_{21} = \beta^*_{31}$.

2.2. Unconstrained parametrization of feasible models

A feasible model can be defined by a linear space of parameters

$$\mathcal{L}_M = \{\boldsymbol{\beta} \in \mathbb{R}^p : \mathbf{A}_{0M}\boldsymbol{\beta} = 0\}, \quad (5)$$

where $\mathbf{A}_{0M}$ is a $(p-q) \times p$ matrix encoding q elementary constraints induced by the model. Such a constraint matrix can be expressed in many ways. In particular, every linear space can be spanned by different vectors. The number of such vectors can be greater than the dimension of the space when they are linearly dependent. In order to unify the form of a constraint matrix, we introduce the notion of regular form, which is described in the Appendix A. We assume that $\mathbf{A}_{0M}$ is in regular form. Let $\mathbf{A}_{1M}$ be a $q \times p$ complement of $\mathbf{A}_{0M}$ to invertible matrix A_M, that is:

$$\mathbf{A}_M = \left[\frac{\mathbf{A}_{1M}}{\mathbf{A}_{0M}}\right].$$

Denote:

$$\mathbf{A}_M^{-1} = \left[\begin{array}{c|c} \mathbf{A}_M^1 & \mathbf{A}_M^0 \end{array}\right], \quad (6)$$

where $\mathbf{A}_M^1$ is a $p \times q$ matrix. In order to replace a constrained by an unconstrained parametrization change of variables in model M is performed. Let $\boldsymbol{\beta}_M \in \mathcal{L}_M$ and $\boldsymbol{\xi}_M = \mathbf{A}_{1M}\boldsymbol{\beta}_M$. We have:

$$\boldsymbol{\beta}_M = \mathbf{A}_M^1 \boldsymbol{\xi}_M. \quad (7)$$

Indeed,

$$\boldsymbol{\beta}_M = \mathbf{A}_M^{-1}\mathbf{A}_M\boldsymbol{\beta}_M = \mathbf{A}_M^{-1}\left[\frac{\mathbf{A}_{1M}\boldsymbol{\beta}_M}{\mathbf{A}_{0M}\boldsymbol{\beta}_M}\right] = \left[\begin{array}{c|c} \mathbf{A}_M^1 & \mathbf{A}_M^0 \end{array}\right]\left[\frac{\boldsymbol{\xi}_M}{\mathbf{0}}\right] = \mathbf{A}_M^1\boldsymbol{\xi}_M.$$

From equation (7) we obtain $\mathbf{X}\boldsymbol{\beta}_M = \mathbf{Z}_{1M}\boldsymbol{\xi}_M$, where $\mathbf{Z}_{1M} = \mathbf{X}\mathbf{A}_M^1$ and $\mathcal{L}_M = \{\mathbf{A}_M^1\boldsymbol{\xi} : \boldsymbol{\xi} \in \mathbb{R}^q\}$. Let us notice that $\mathcal{L}_M$ is a linear space spanned by columns of $\mathbf{A}^1_M$. The dimension of space $\mathcal{L}_M$ will be called the size of model M and denoted by $|M|$. Note that $|M| = q$.

Example 1 continued. Matrices $\mathbf{A}_M, \mathbf{A}^1_M, \mathbf{Z}_{1M}$ and $\boldsymbol{\xi}_M$ are:

$$\mathbf{A}_M = \left[\begin{array}{ccccc} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ \hline 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{array}\right], \ \mathbf{A}^1_M = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \ \mathbf{Z}_{1M} = \begin{bmatrix} 1 & -0.96 & 0 \\ 1 & -0.29 & 0 \\ 1 & 0.26 & 1 \\ 1 & -1.15 & 1 \\ 1 & 0.2 & 1 \\ 1 & 0.03 & 1 \\ 1 & 0.09 & 0 \\ 1 & 1.12 & 0 \end{bmatrix},$$

$$\boldsymbol{\xi}_M = (\xi_1, \xi_2, \xi_3)^T \ , \ \xi_1 = \beta^*_{00}, \ \xi_2 = \beta^*_{10} \ , \ \xi_3 = \beta^*_{21} = \beta^*_{31}.$$

One can see that a change from a constrained to an unconstrained problem was done by adding and deleting columns of the model matrix.

The OLS estimator of $\boldsymbol{\beta}^*$ constrained to $\mathcal{L}_M$ is given by the following expression:

$$\widehat{\boldsymbol{\beta}}_M = \mathbf{A}^1_M \widehat{\boldsymbol{\xi}}_M, \ \text{where } \widehat{\boldsymbol{\xi}}_M = \left(\mathbf{Z}^T_{1M}\mathbf{Z}_{1M}\right)^{-1} \mathbf{Z}^T_{1M}\mathbf{y}. \tag{8}$$

Note that $\mathbf{A}_{0M}\widehat{\boldsymbol{\beta}}_M = \mathbf{A}_{0M}\mathbf{A}^1_M\widehat{\boldsymbol{\xi}}_M = 0$ and thus indeed $\widehat{\boldsymbol{\beta}}_M \in \mathcal{L}_M$. We define the inclusion relation between two models M_1 and M_2 by inclusion of linear spaces

$$M_1 \subseteq M_2 \text{ denotes } \mathcal{L}_{M_1} \subseteq \mathcal{L}_{M_2} \tag{9}$$

and intersection of two models M_1 and M_2 by intersection of linear spaces:

$$M_1 \cap M_2 \text{ as a model defined by } \mathcal{L}_{M_1} \cap \mathcal{L}_{M_2}. \tag{10}$$

A feasible model M will be called a true model if $\boldsymbol{\beta}^* \in \mathcal{L}_M$. A true model with minimal size will be denoted by T. Observe that T is unique because $\mathbf{X}$ is a full rank matrix.

Example 1 continued. For the illustrative example the true model T is $T = (\{1\}, \{\{1,4\}, \{2,3\}\})$. The dimensions of the considered models are $|F| = p = 5$, $|T| = 3$.

2.3. Residual sum of squares and bayesian information criterion for feasible models

Let $\mathbf{H}_M = \mathbf{Z}_{1M}\left(\mathbf{Z}^T_{1M}\mathbf{Z}_{1M}\right)^{-1}\mathbf{Z}^T_{1M}$. Observe that $\mathbf{H}_M\mathbf{X}\boldsymbol{\beta}^* = \mathbf{X}\boldsymbol{\beta}^*$ for $M \supseteq T$. We define residual sum of squares for model M as $RSS_M = \|\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_M\|^2$. From equation (8) we have:

$$RSS_M = \|\mathbf{y} - \mathbf{Z}_{1M}\widehat{\boldsymbol{\xi}}_M\|^2 = \|(\mathbb{I} - \mathbf{H}_M)\mathbf{y}\|^2.$$

Let us denote:

$$\Delta_M = \boldsymbol{\beta}^{*T}\mathbf{X}^T(\mathbb{I} - \mathbf{H}_M)\mathbf{X}\boldsymbol{\beta}^* = \|\mathbf{X}\boldsymbol{\beta}^* - \mathbf{X}\boldsymbol{\beta}^*_M\|^2, \tag{11}$$

where $\boldsymbol{\beta}^*_M = \arg\min_{\boldsymbol{\beta}\in\mathcal{L}_M} \|\mathbf{X}\boldsymbol{\beta}^* - \mathbf{X}\boldsymbol{\beta}\|^2$. Notice that $\widehat{\boldsymbol{\beta}}_M \xrightarrow{P} \boldsymbol{\beta}^*_M$ with $n \to \infty$. The following decomposition of RSS in linear models is trivial, hence we omit the proof:

Proposition 1.

$$RSS_M = \Delta_M + 2\boldsymbol{\beta}^{*T}\mathbf{X}^T(\mathbb{I} - \mathbf{H}_M)\boldsymbol{\epsilon} + \boldsymbol{\epsilon}^T(\mathbb{I} - \mathbf{H}_M)\boldsymbol{\epsilon}.$$

In particular for $M \supseteq T$

$$RSS_M = \boldsymbol{\epsilon}^T(\mathbb{I} - \mathbf{H}_M)\boldsymbol{\epsilon} \sim \sigma^2\chi^2_{n-|M|}.$$

Therefore, the predictions for a constrained problem can be obtained through projecting the observations on the space spanned by columns of the model matrix for the equivalent unconstrained problem. Hence, decompositions and asymptotic properties of residual sums of squares for feasible models are inherited from the unconstrained linear models.

Bayes Information Criterion for model M is defined as:

$$BIC_M = n \log RSS_M + \log(n)|M|.$$

The goal of our method is to find the best feasible model according to BIC, taking into account that the number of feasible models grows exponentially with p. Since for the k-th factor the 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 a greedy backward search.

3. DMR algorithm

In this section we introduce the DMR algorithm. Because of troublesome notations, in order to make the description of the algorithm more intuitive, we present here a general idea of the algorithm. In particular, we give the details of step 3 of the algorithm in the Appendix B.

Assuming that $\mathbf{X}$ is of full rank the QR decomposition of the model matrix is $\mathbf{X} = \mathbf{QR}$, where $\mathbf{Q}$ is $n \times p$ orthogonal matrix and $\mathbf{R}$ is $p \times p$ upper triangular matrix. Denote the minimum variance unbiased estimators of $\boldsymbol{\beta}$ and σ^2 for the full model F as:

$$\widehat{\boldsymbol{\beta}} = \mathbf{R}^{-1}\mathbf{z} \text{ and } \widehat{\sigma}^2 = \frac{\|\mathbf{y}\|^2 - \|\mathbf{z}\|^2}{n-p}, \text{ where } \mathbf{z} = \mathbf{Q}^T\mathbf{y}. \tag{12}$$

Let us denote

$$\widehat{\boldsymbol{\beta}} = [\widehat{\beta}_{jk}]_{\substack{j \in N_k \\ k \in N}}, \ \mathbf{R}^{-1} = [r_{jk,st}]_{\substack{j \in N_k \\ s \in N_t \\ k,\, t \in N}},$$

then

$$\widehat{\beta}_{jk} = \mathbf{r}_{jk}^T \mathbf{z}, \text{ where } j \in N_k, k \in N$$

and $\mathbf{r}_{jk}$ is a row of $\mathbf{R}^{-1}$.

Algorithm 1 DMR (Delete or Merge Regressors)

Input: $\mathbf{y}$, $\mathbf{X}$

1. Computation of t-statistics

Compute the QR decomposition of the full model matrix, obtaining matrix $\mathbf{R}^{-1}$, vector $\mathbf{z}$ and variance estimator $\widehat{\sigma}^2$ as in equation (12). Calculate squared t-statistics:

1. for all elementary constraints defined in (2):

$$t_{1jk}^2 = \frac{\widehat{\beta}_{jk}^2}{\widehat{Var}(\widehat{\beta}_{jk})} = \frac{(\mathbf{r}_{jk}^T \mathbf{z})^2}{\widehat{\sigma}^2 \|\mathbf{r}_{jk}\|^2} \quad \text{for} \quad j \in N_k \setminus \{0\}, \;\; k \in N,$$

2. for all elementary constraints defined in (3):

$$t_{ijk}^2 = \frac{(\widehat{\beta}_{ik} - \widehat{\beta}_{jk})^2}{\widehat{Var}(\widehat{\beta}_{ik} - \widehat{\beta}_{jk})} = \frac{((\mathbf{r}_{ik} - \mathbf{r}_{jk})^T z)^2}{\widehat{\sigma}^2 \|\mathbf{r}_{ik} - \mathbf{r}_{jk}\|^2}$$

   for $i, j \in N_k$, $i \neq j$, $k \in N \setminus \{0\}$.

2. Agglomerative clustering for factors (using complete linkage clustering)

For each factor perform agglomerative clustering using $\mathbf{D}_k = \left[d_{ijk}\right]_{ij}$ as dissimilarity matrix for $k \in N \setminus \{0\}$:

1. $d_{1jk} = d_{j1k} = t_{1jk}^2$ for $j \in N_k$,
2. $d_{ijk} = t_{ijk}^2$ for $i, j \in N_k$, $i \neq j$,
3. $d_{iik} = 0$ for $i \in N_k$.

We denote cutting heights obtained from the clusterings as $\mathbf{h}_1^T, \mathbf{h}_2^T, \ldots, \mathbf{h}_l^T$.

3. Sorting constraints (hypotheses) according to the squared t-statistics

Combine vectors of cutting heights: $\mathbf{h} = [0, \mathbf{h}_0^T, \mathbf{h}_1^T, \ldots, \mathbf{h}_l^T]^T$, where $\mathbf{h}_0$ is a vector of squared t-statistics for constraints concerning continuous variables and 0 corresponds to the full model. Sort elements of $\mathbf{h}$ in increasing order and construct a corresponding $(p-1) \times p$ matrix $\mathbf{A}_0$ of consecutive constraints.

4. Computation of RSS using a recursive formula in a nested family of models

Perform QR decomposition of the matrix $\mathbf{R}^{-T}\mathbf{A}_0^T$ obtaining the orthogonal matrix $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_{p-1}]$. Set $\text{RSS}_{M_0} = \|\mathbf{y}\|^2 - \|\mathbf{z}\|^2$ for a model without constraints. For $m = 1, \ldots, p-1$

$$\text{RSS}_{M_m} = \text{RSS}_{M_{m-1}} + (\mathbf{w}_m^T \mathbf{z})^2,$$

where M_m denotes a model with constraints defined by m first rows of $\mathbf{A}_0$. The last formula is derived in the Appendix C, see equation (22).

5. Choosing the best model according to BIC

Calculate

$$\text{BIC}_{M_m} = n \log \text{RSS}_{M_m} + (p-m)\log(n)$$

for $m = 0, \ldots, p-1$. Selected model $\widehat{T}$ is the model minimizing BIC among models on the nested path:

$$\widehat{T} = \underset{\substack{m \\ 0 \le m \le p-1}}{\arg\min} \; \text{BIC}_{M_m}.$$

Output: $\widehat{T}$

The time complexities of successive steps of the DMR algorithm are $O(np^2)$ for QR decomposition in step 1, $O(p^2)$ for hierarchical clustering in step 2, $O(p^3)$

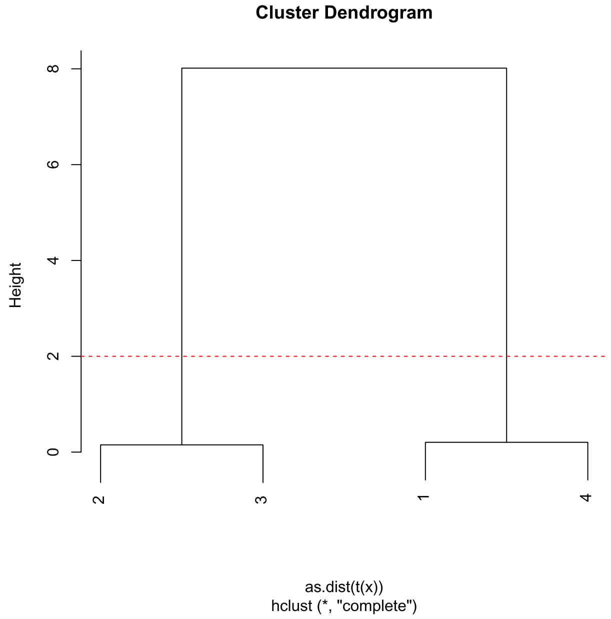

FIG 2. *Dendrogram for Example 1.*

for QR decomposition used in step 4. The dominating operation in the described procedure is the QR decomposition of the full model matrix. Hence, the overall time complexity of the DMR algorithm is $O(np^2)$.

Example 1 continued. For the illustrative example we have:

$$t^2_{110} = 9.35\ ,\ \mathbf{D_1} = \begin{bmatrix} 0 & t^2_{121} & t^2_{131} & t^2_{141} \\ t^2_{121} & 0 & t^2_{231} & t^2_{241} \\ t^2_{131} & t^2_{231} & 0 & t^2_{341} \\ t^2_{141} & t^2_{241} & t^2_{341} & 0 \end{bmatrix} = \begin{bmatrix} 0 & 8.01 & 4.52 & 0.20 \\ 8.01 & 0 & 0.15 & 3.09 \\ 4.52 & 0.15 & 0 & 2.91 \\ 0.20 & 3.09 & 2.91 & 0 \end{bmatrix},$$

$$\mathbf{h} = [0, 0.15, 0.20, 8.01, 9.33]^T\ ,\ \mathbf{A_0} = \begin{array}{c} \begin{array}{ccccc} \beta_{00} & \beta_{10} & \beta_{21} & \beta_{31} & \beta_{41} \end{array} \\ \begin{bmatrix} 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix} \end{array},$$

$$\mathbf{BIC} = [28.33, 26.65, 25.36, 34.68, 39.59]^T.$$

Observe that the selected model $\widehat{T}$ is the true model T. The dendrogram and cutting heights for the illustrative example obtained from clustering in step 2 are shown in Figure 2. The horizontal dashed line corresponds to the optimal partition chosen by BIC.

4. Asymptotic properties of the DMR algorithm

In Algorithm 1 and all the simulations and examples we assumed complete linkage in hierarchical clustering and BIC for selection in the nested family of models. The proof of consistency is more general: the linkage criterion has to be a convex combination of the minimum and maximum of the pairwise distances between clusters (see equation 24 in Appendix D) and generalized information criterion is used for final model selection:

$$GIC_M = n \log RSS_M + r_n|M|,$$

where r_n is the penalty for model size. Note that well known criteria AIC and BIC are special cases of GIC, if $r_n = 2$ and $r_n = \log(n)$, respectively.

In this section we use $f_n \prec g_n$ to denote $f_n = o(g_n)$. We allow the number of predictors p_n to grow monotonically 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. Uniquely defined model T, which is fixed and does not depend on the 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.
2. A set $\mathcal{M}_\mathcal{V}$ of models with one constraint imposed which is false:
$$\mathcal{M}_\mathcal{V} = \{M \subseteq F : |M| = |F| - 1 \text{ and } T \nsubseteq M\},$$
3. A set $\mathcal{M}_\mathcal{T}$ of models with one constraint imposed which is true:
$$\mathcal{M}_\mathcal{T} = \{M \subseteq F : |M| = |F| - 1 \text{ and } T \subseteq M\}.$$

We denote:

$$\Delta = \min_{M \in \mathcal{M}_\mathcal{V}} \Delta_M, \tag{13}$$

where Δ_M was defined in equation (11). Let us notice that from equation (8) we get

$$\text{Var}\left(\widehat{\boldsymbol{\beta}}_M\right) = \mathbf{A}_M^1 \text{Var}\left(\widehat{\boldsymbol{\xi}}_M\right) \mathbf{A}_M^{1T} = \mathbf{A}_M^1 \left(\mathbf{A}_M^{1T}\mathbf{X}^T\mathbf{X}\mathbf{A}_M^1\right)^{-1} \mathbf{A}_M^{1T}.$$

Then

$$\text{Var}\left(\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}^*\right)\right) = n\mathbf{A}_M^1 \left(\mathbf{A}_M^{1T}\mathbf{X}^T\mathbf{X}\mathbf{A}_M^1\right)^{-1} \mathbf{A}_M^{1T}.$$

Additionally, for finite p, independent of n, if $\frac{1}{n}\mathbf{X}^T\mathbf{X} \to \boldsymbol{\Sigma} > 0$ then

$$\text{Var}\left(\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}^*\right)\right) \to \boldsymbol{\Sigma}_M = \mathbf{A}_M^1 \left(\mathbf{A}_M^{1T}\boldsymbol{\Sigma}\mathbf{A}_M^1\right)^{-1} \mathbf{A}_M^{1T}.$$

Theorem 1. *Assume that* $\mathbf{X}$ *is of full rank and* $p_n \prec r_n \prec \min(n, \Delta)$. *Let* $\widehat{T}$ *be the model selected by DMR, where linkage criterion for hierarchical clustering is a convex combination of minimum and maximum of the pairwise distances between clusters. Then* *(a)* $\lim_{n\to\infty} \mathbb{P}(\widehat{T} = T) = 1$,

(b) $\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{\widehat{T}} - \boldsymbol{\beta}^*\right) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{\Sigma}_T)$ *if additionally* p *is finite, independent of* n *and* $\frac{1}{n}\mathbf{X}^T\mathbf{X} \to \boldsymbol{\Sigma} > 0$.

Proof can be found in the Appendix D.

5. Numerical experiments

All experiments were performed using functions implemented in `R` package called `DMR`, which is available at the `CRAN` repository. The main function in the package is called `DMR` and implements the DMR algorithm with an optional method of hierarchical clustering (default is complete) and a value of r_n in GIC (default is $\log(n)$). The package also contains other functions that are modifications of the DMR algorithm, such as stepDMR which assumes recalculation of t-statistics after accepting every new elementary constraint and DMR4glm which can be used for model selection in generalized linear models.

We compared 2 groups of algorithms. The first one contains 3 stepwise procedures stepBIC, ffs BIC and DMR. The second group are 2 Lasso-based methods: CAS-ANOVA and gvcm. Procedure stepBIC is implemented in the function `stepAIC` in `R` package `MASS` and does not perform factor partitions but either deletes or keeps any of categorical predictors. A factor forward stepwise procedure (ffs BIC), implemented in `R` package `gvcm.cat` is similar to DMR but differs in the search direction (DMR is backward and ffs BIC is forward) and in the criterion of selection of the best step (DMR uses t-statistics calculated only once and hierarchical clustering and ffs BIC recalculates criterion in every step). For DMR the complete linkage method of clustering and BIC were used. Algorithm gvcm is implemented in `R` package `gvcm.cat` where by default there are no adaptive weights and crossvalidation is used for choosing the λ parameter. We used adaptive weights and BIC criterion for choosing the tuning parameter since we got better results then. Implementation of CAS-ANOVA can be found on the website `http://www4.stat.ncsu.edu/~bondell/Software/CasANOVA/CasANOVA.R`. Here the default BIC was used for choosing the λ parameter making all the methods dependent on the same criterion of choosing the tuning parameters. Adaptive weights are also default in CAS-ANOVA. When using the two Lasso-based algorithms we found difficult the selection of the λ grid. In all the experiments we tried different grids: the default ones and ours both on linear and logarithmic scales presenting only the best results.

We describe three simulation experiments. In Section 5.2 results regarding an experiment constructed in the same way as in Bondell and Reich (2009) is presented. The model 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 5.3. In Section 5.4 we summarize the results of an experiment regarding generalized linear models. In this experiment only 4 algorithms were compared since CAS-ANOVA applies only to normal distribution.

In Section 5.1 we introduce measures of performance which are generalizations of popular true positive rate and false discovery rate on categorical predictors. We call them TPR^* and FDR^*. In comparison to generalizations introduced in Gertheiss and Tutz (2010) and Bondell and Reich (2009), which we call TPR and FDR, our measures don't diminish the influence of continuous predictors and factors with a small number of levels. Hence, for evaluation of the model selection methods we used following criteria: true model (TM) represents the percentage of times the procedure chose the entirely correct model. Correct factors (CF) represents the percentage of times the non-significant factors were eliminated and the true factor was kept. 1−TPR, FDR, 1−TPR* and FDR* are averaged errors made by selectors described in Section 5.1. 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 last Section 5.5 refers to two real data examples where barley yield and prices of apartments in Munich were modeled.

5.1. Measures of performance

When performing simulations, results are usually compared to the underlying truth. Traditionally, for model selection with only continuous predictors measures such as true positive rate (TPR) or false discovery rate (FDR) are used. In the literature (Gertheiss and Tutz (2010), Bondell and Reich (2009)) their generalization to both continuous and categorical predictors can be found.

Let us consider sets of elementary constraints corresponding to the true and selected models determined by sets of indexes:

$$\mathcal{B} = \{(i,j,k): \; i \neq j, i,j \in N_k, k \in N \setminus \{0\}, \; (\boldsymbol{\beta}^*)_{ik} = (\boldsymbol{\beta}^*)_{jk}\}$$
$$\cup \{(j,k) : j \in N_k, k \in N, (\boldsymbol{\beta}^*)_{jk} = 0\}$$

and

$$\widehat{\mathcal{B}} = \{(i,j,k): \; i \neq j, i,j \in N_k, k \in N \setminus \{0\}, \; (\widehat{\boldsymbol{\beta}}_{\widehat{T}})_{ik} = (\widehat{\boldsymbol{\beta}}_{\widehat{T}})_{jk}\}$$
$$\cup \{(j,k) : j \in N_k, k \in N, (\widehat{\boldsymbol{\beta}}_{\widehat{T}})_{jk} = 0\}.$$

True positive rate is the proportion of true differences which were correctly identified to all true differences, meaning ratio of the number of true elementary constraints which were found by the selector to the number of all true elementary constraints $TPR = |\mathcal{B} \cap \widehat{\mathcal{B}}|/|\mathcal{B}|$. False discovery rate is the proportion of false differences which were classified as true to all differences classified as true, meaning ratio of the number of false elementary constraints which were accepted by the selector to the number of all accepted elementary constraints $FDR = 1 - |\mathcal{B} \cap \widehat{\mathcal{B}}|/|\widehat{\mathcal{B}}|$.

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 continuous predictors is 5 and the number of possible elementary constraints equals 5. The number of parameters for the categorical variable is also 5, whereas the number of possible elementary constraints is $\binom{5}{2} = 10$.

We introduce a different generalization of the traditional performance measures using dimensions of linear spaces which define the true and selected models. We consider two models: true model T and selected model $\widehat{T}$.

We define true positive rate coefficient as $TPR^* = |T \cap \widehat{T}|/|T|$ and false discovery rate coefficient as $FDR^* = 1 - |T \cap \widehat{T}|/|\widehat{T}|$, where $T \cap \widehat{T}$ is defined according to equation (10). This generalization is more fair since the influence of every parameter on the coefficients is equal. In the article the attention is focused on values: $1 - TPR^*$ and FDR^*, which correspond to the errors made by selector.

5.2. Experiment 1

The layout of this experiment is the same as in Bondell and Reich (2009). Despite using different λ grids, we weren't able to obtain as good results for CAS-ANOVA as in the original paper. However, the results for DMR are much better in terms of TM than those for CAS-ANOVA originally reported in Bondell and Reich (2009). The experimental model consists of three factors having eight, four and three levels, respectively. The true model is $T = (P_1, P_2, P_3)$, where

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

The response $\mathbf{y}$ was generated using the true model:

$$\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbb{I}),$$

where

$$\begin{aligned}\boldsymbol{\mu} &= \mathbf{1}_n\beta_{00}^* + \mathbf{X}_1\boldsymbol{\beta}_1^* + \mathbf{X}_2\boldsymbol{\beta}_2^* + \mathbf{X}_3\boldsymbol{\beta}_3^* \\ &= \mathbf{1}_n \cdot 2 + \mathbf{X}_1(0,-3,-3,-3,-3,-2,-2)^T + \mathbf{X}_2(0,0,0)^T + \mathbf{X}_3(0,0)^T.\end{aligned}$$

A balanced design was used with c observations for each combination of factor levels, which gives $n = 96 \cdot c$, $c = 1,2,4$.

The data was generated 1000 times. The best results for $\lambda_{\text{CAS-ANOVA}} = (0.1, 0.2, \ldots, 3)^T$ and $\lambda_{\text{gvcm}} = (0.01, 0.02, \ldots, 3)^T$ together with outcomes from other methods are summarized in Table 1. The results of Experiment 1 indicate that DMR and ffs BIC algorithms performed almost twice better than CAS-ANOVA and gvcm in terms of choosing the true model. Our procedure and ffs BIC chose approximately smaller models with dimension closer to the dimension of the underlying true model, whose number of parameters is three. There were no significant differences between mean squared errors of prediction for all considered algorithms. The main conclusion, that DMR and ffs BIC procedures choose models which are smaller and closer to the proper one, is supported by the obtained values of 1 - TPR* and FDR*.

TABLE 1
Results of the simulation study, Experiment 1

n	Algorithm	TM(%)	CF(%)	1-TPR	FDR	1-TPR*	FDR*	MSEP±sd	MD±sd
96	DMR	44	73	0.05	0.09	0.1	0.19	1.091±.179	3.4±.7
	ffs BIC	42	73	0.04	0.09	0.1	0.2	1.091±.179	3.5±.7
	CAS-ANOVA	17	83	0.04	0.14	0.06	0.33	1.104±.175	5.5± 1.7
	gvcm	11	49	0.08	0.15	0.1	0.34	1.118±.179	4.5±1.6
	stepBIC	0	97	0	0.29	0	0.63	1.089±.171	8.1±.4
192	DMR	66	82	0.01	0.05	0.02	0.1	1.036±.11	3.3±.6
	ffs BIC	67	83	0.01	0.05	0.02	0.1	1.035±.11	3.3±.5
	CAS-ANOVA	33	93	0	0.09	0.01	0.24	1.049±.109	4.9±1.3
	gvcm	27	60	0.01	0.11	0.02	0.27	1.049±.11	4.3±1.2
	stepBIC	0	99	0	0.29	0	0.63	1.046±.109	8±.2
384	DMR	80	89	0	0.03	0	0.05	1.013±.074	3.2±.4
	ffs BIC	79	89	0	0.03	0	0.05	1.013±.074	3.2±.4
	CAS-ANOVA	50	97	0	0.06	0	0.17	1.022±.074	4.2±1.2
	gvcm	49	77	0	0.06	0	0.16	1.02±.074	3.8±1
	stepBIC	0	100	0	0.29	0	0.63	1.022±.074	8±.1

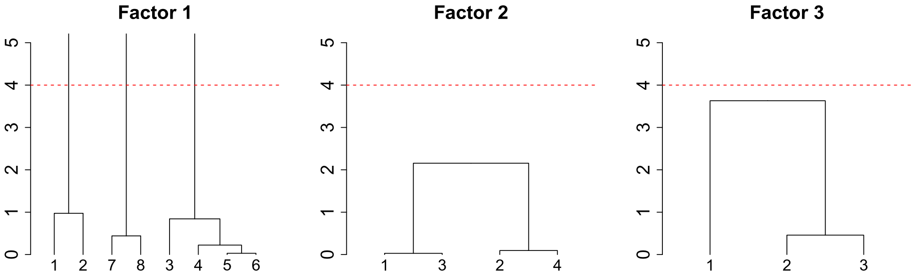

FIG 3. *An examplary run of the DMR algorithm for Experiment 1.*

TABLE 2
Computation times divided by the computation time of `lm.fit`, results obtained using `system.time` function

c	n	DMR	ffs BIC	CASANOVA	gvcm	stepBIC
1	96	87	883	234	250	71
4	384	36	526	89	245	31
20	1920	19	394	21	739	16

An exemplary run of the DMR algorithm is shown in Figure 3. The horizontal dotted line indicates the cutting height for the best model chosen by BIC.

In Table 2 the computation times of the 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 and gvcm are given for only one value of λ. By default, the searched lambda grid is of length 50 and 5001, respectively. One can see that DMR is significantly faster than ffs BIC, CAS-ANOVA and gvcm.

TABLE 3

Results of the simulation study, Experiment 2

n	Algorithm	TM(%)	1-TPR	FDR	1-TPR*	FDR*	MSEP±sd	MD±sd
128	DMR	68	0	0.03	0	0.05	1.076±.148	7.4±.6
	ffs BIC	60	0.01	0.04	0.01	0.06	1.081±.15	7.3±.8
	CAS-ANOVA	17	0	0.13	0	0.21	1.11±.153	9.9±1.6
	gvcm	12	0.02	0.11	0.01	0.23	1.113±.154	8.2±1.5
	stepBIC	0	0	0.25	0	0.42	1.101±.148	12.1±.4
256	DMR	78	0	0.02	0	0.03	1.033±.093	7.2±.5
	ffs BIC	54	0	0.03	0	0.07	1.034±.093	7.4±.8
	CAS-ANOVA	27	0	0.1	0	0.16	1.049± .096	9.2±1.4
	gvcm	24	0	0.07	0	0.17	1.047±.096	7.5±1.3
	stepBIC	0	0	0.25	0	0.42	1.049±.095	12.1±.3
512	DMR	88	0	0.01	0	0.02	1.015±.066	7.1±.4
	ffs BIC	85	0	0.01	0	0.02	1.016±.066	6.9±.6
	CAS-ANOVA	46	0	0.06	0	0.1	1.024±.067	8.4±1.2
	gvcm	35	0	0.05	0	0.12	1.021±.067	7±1.1
	stepBIC	0	0	0.25	0	0.42	1.023±.067	12±.2

5.3. Experiment 2

In the second experiment a model containing not only categorical predictors, but also continuous variables is considered. The response $\mathbf{y}$ was generated from the model with one factor with eight levels and eight continuous variables:

$$\begin{aligned}\mathbf{y} &= \mathbf{V}_0\boldsymbol{\alpha}_0 + \mathbf{V}_1\boldsymbol{\alpha}_1 + \boldsymbol{\varepsilon}\\ &= \mathbf{V}_0(1,0,1,0,1,0,1,0)^T + \mathbf{V}_1(0,0,-2,-2,-2,-2,4,4)^T + \boldsymbol{\varepsilon},\end{aligned}$$

where $\mathbf{V}_0$ was generated from the multivariate normal distribution with autoregressive correlation structure with $\rho = 0.8$. The first $2\cdot 16\cdot c$ rows were generated using mean vector $(1,1,0,0,0,0,0,0)^T$, then $4\cdot 16\cdot c$ observations using mean vector $(0,0,1,1,1,1,0,0)^T$ and the last $2\cdot 16\cdot c$ observations using mean vector $(0,0,0,0,0,0,1,1)^T$, according to the underlying true partition of the factor. $c = 1,2,4$, hence $n = 128\cdot c$. $\mathbf{V}_1$ is a matrix of dummy variables encoding levels of the factor and $\boldsymbol{\varepsilon}$ was generated from zero-mean normal distribution, $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0},\mathbb{I})$. The data was generated 1000 times.

The best results for $\lambda_{\text{CAS-ANOVA}} = (0.1,0.2,\ldots,3)^T$ and $\lambda_{\text{gvcm}} = (0.01,0.02,\ldots,5)^T$ together with outcomes from other methods are summarized in Table 3. Despite the fact that additional continuous variables were correlated, the obtained results show a considerable advantage of the DMR algorithm over other methods.

5.4. Experiment 3

Simultaneous deleting continuous variables and merging levels of factors can also be considered in the framework of generalized linear models. The problem has already been discussed in Oelker, Gertheiss and Tutz (2014), where L_1 regularization was used. After replacing squared t-statistics with squared

TABLE 4
Results of the simulation study for logistic regression, Experiment 3

n	Algorithm	TM	CF	1-TPR	FDR	1-TPR*	FDR*	MSEP±sd	MD±sd
96	DMR	6	62	0.21	0.15	0.38	0.35	0.304±.049	3.1±1.2
	ffs BIC	7	72	0.21	0.14	0.37	0.35	0.302±.049	3.1±.8
	gvcm	0	21	0.18	0.32	0.27	0.61	0.317±.062	6.4±2.9
	stepBIC	0	96	0.00	0.29	0.00	0.63	0.299±.049	8±.6
192	DMR	25	81	0.16	0.09	0.25	0.23	0.296±.036	3±.7
	ffs BIC	21	82	0.17	0.10	0.28	0.26	0.293±.034	3±.7
	gvcm	1	26	0.15	0.26	0.19	0.52	0.296±.038	5.8±2.6
	stepBIC	0	99	0.00	0.29	0.00	0.63	0.291±.034	8±.2
384	DMR	55	88	0.06	0.06	0.12	0.14	0.29±.023	3.1±.5
	ffs BIC	51	88	0.06	0.06	0.12	0.16	0.29±.023	3.2±.5
	gvcm	6	37	0.08	0.20	0.10	0.43	0.289±.022	5.5±2.5
	stepBIC	0	100	0.00	0.29	0.00	0.63	0.289±.022	8±.2
768	DMR	79	92	0.01	0.03	0.03	0.07	0.29±.016	3.1±.4
	ffs BIC	79	92	0.01	0.03	0.03	0.06	0.29±.016	3.1±.4
	gvcm	20	48	0.01	0.16	0.02	0.36	0.289±.016	5.2±2.2
	stepBIC	0	100	0.00	0.29	0.00	0.63	0.29±.016	8±.1

TABLE 5
Computation times divided by the computation time of `glm.fit`, results obtained using `system.time` function

c	n	DMR	ffs BIC	gvcm	stepBIC
1	96	103	399	101	40
4	384	68	398	74	28
20	1920	49	377	101	23

Wald's statistics, the DMR algorithm can be easily modified to generalized linear models. Simulation results for the DMR algorithm for logistic regression are presented below. Let us consider a logistic regression model whose linear part consists of three factors defined as in Experiment 1. The response $\mathbf{y}$ was independently sampled from binomial distribution:

$$y_i \sim B\left(1, \frac{\exp(\mu_i)}{1+\exp(\mu_i)}\right),\ i = 1, \dots, n,$$

where μ_i are elements of $\boldsymbol{\mu}$ defined as in Experiment 1, $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)^T$ and $n = 96 \cdot c$ for $c = 1, 2, 4, 8$.

The results of the experiment are summarized in Table 4. The best outcomes for gvcm, presented in the table, were obtained for λ grids $\lambda_{\text{gvcm}} = (0.01, 0.02, \dots, 5)^T$. Again, DMR and ffs BIC show considerable advantage over other model selection methods.

In Table 5 the computation times of the algorithms are summarized. All values are divided by the computation time of `glm.fit` function. The results for gvcm are given for only one value of λ, while by default the searched lambda grid is of length 5001. DMR is again significantly faster than ffs BIC and gvcm.

Table 6
Characteristics of the chosen models for Barley data set

algorithm	model dim	R^2	adj. R^2	BIC
full model	11	.68	.61	416
stepBIC	11	.68	.61	416
CAS-ANOVA λ_2	9	.66	.61	411
gvcm λ_2	7	.66	.6	403
CAS-ANOVA λ_1	6	.61	.58	407
ffs BIC	5	.64	.61	399
DMR	5	.64	.61	399

5.5. Real data examples

Example 1: Barley The data set `barley` from `R` library `lattice` has already been discussed in the literature, for example in Bondell and Reich (2009). The response is the barley yield for each of 5 varieties (Svansota, Manchuria, Velvet, Peatland and Trebi) at 6 experimental farms in Minnesota for each year of the years 1931 and 1932 giving a total of 60 observations. The characteristics of the chosen models using different algorithms are presented in Table 6. The results for the full model which is least squares estimator with all variables were given as a benchmark. For the two Lasso-based algorithms we find difficult the selection of the λ grid. Therefore, the results for CAS-ANOVA are given for two different grids: the first one chosen so that the chosen model was the same as the one described in Bondell and Reich (2009), $\lambda_1 = (25, 25.01, 25.02, \ldots, 35)^T$, and the second wider superset of the first one, $\lambda_1 = (0.1, 0.2, 0.3, \ldots, 35)^T$. We used λ_2 grid also for gvcm.

The results show that stepwise methods give smaller models with smaller BIC values than the Lasso-based methods. The additional advantage of DMR and ffs BIC is lack of a troublesome tuning parameter.

Example 2: Miete The data set `miete03` comes from `http://www.statistik.lmu.de/service/datenarchiv`. The data consists of 2053 households interviewed for the Munich rent standard 2003. The response is monthly rent per square meter in Euros. 8 categorical and 3 continuous variables give 36 and 4 (including the intercept) parameters. The data is described in detail in Gertheiss and Tutz (2010).

Model selection was performed using five methods: DMR, ffs BIC, CAS-ANOVA, gvcm and stepBIC. Characteristics of the chosen models are shown in Table 7 with results for the full model added for comparison.

The reason of the lack of results for ffs BIC in the part of Table 7 is that the algorithm required to allocate too much memory (factor urban district has 25 levels).

We can conclude that DMR procedure and ffs BIC chose much better models than other compared methods in terms of BIC. However, DMR method can be applied to problems with larger number of parameters.

Table 7
Characteristics of the chosen models for Miete data set

Selection method	Model dimension	R^2	adj.R^2	BIC
Full model	40	.94	.94	23037
CAS-ANOVA	31	.94	.94	22972
gvcm	26	.94	.94	22933
DMR	12	.94	.94	22833
stepBIC	11	.94	.94	22847

6. Discussion

We propose the DMR method which combines deleting continuous variables and merging levels of factors in linear models. DMR relies on ordering of elementary constraints using squared t-statistics and choosing the best model according to BIC in the nested family of models. A slightly modified version of the DMR algorithm can be applied to generalized linear models.

We proved that DMR is a consistent model selection method. The main advantage of our theorem over the analogous one for the Lasso based methods (CAS-ANOVA, gvcm) is that we allow the number of predictors to grow to infinity.

We show in simulations that DMR and ffs BIC are more accurate than the Lasso-based methods. However, DMR is much faster and less memory demanding in comparison to ffs BIC. Our results are not exceptional in comparison to others in the literature. In Example 1 in Zou and Li (2008) a similar simulation setup to our Experiment 1, $n = 96$, has been considered. The adaptive Lasso method (denoted there as one-step LOG) was outperformed by exhaustive BIC with 66 to 73 percent of true model selection accuracy. We repeated the simulations and got similar results with 76 percent for the Zheng-Loh algorithm (described in Zheng and Loh (1995)), which is DMR with just continuous variables. Thus, in the Zou and Li experiment the advantage of the Zheng-Loh algorithm over the adaptive Lasso is not as large as in our work, but Zou and Li used a better local linear approximations (LLA) of the penalty function in the adaptive Lasso implementation. Recall that both CAS-ANOVA and gvcm employ the local quadratic approximation (LQA) of the penalty function.

The superiority of DMR over the Lasso based methods in our experiments not only comes from weakness of LQA used in the adaptive Lasso implementation. Greedy subset selection methods similar to the Zheng-Loh algorithm have been proposed many times. Recently, in Pokarowski and Mielniczuk (2015) a combination of screening of predictors by the Lasso with the Zheng-Loh greedy selection for high-dimensional linear models has been proposed. The authors showed both theoretically and experimentally that such combination is competitive to the Multi-stage Convex Relaxation described in Zhang (2010), which is least squares with capped l_1 penalty implemented via LLA.

Appendix A: Regular form of constraint matrix

We say that $\mathbf{A}_{0M}$ is in regular form if it can be complemented to $\mathbf{A}_M$ so that:

$$\mathbf{A}_M = \left[\begin{array}{c} \mathbf{A}_{1M} \\ \hline \mathbf{A}_{0M} \end{array}\right] = \left[\begin{array}{cc} \mathbb{I} & 0 \\ \hline \mathbf{B}_M & \mathbb{I} \end{array}\right], \tag{14}$$

where $\mathbf{B}_M$ is a matrix consisting of $0, -1, 1$. Then, using Schur complement we get:

$$\mathbf{A}_M^{-1} = \left[\begin{array}{c|c} \mathbb{I} & 0 \\ -\mathbf{B}_M & \mathbb{I} \end{array}\right] = \left[\begin{array}{c|c} \mathbf{A}_M^1 & \mathbf{A}_M^0 \end{array}\right]. \tag{15}$$

Constraint matrix in regular form can always be obtained by a proper permutation of model's parameters. Let us denote clusters in each partition: $P_{Mk} = (C_{Mik})_{i=1}^{i_k}$, where i_k is the number of clusters, $k \in N \setminus \{0\}$ and minimal elements in each cluster as $j_{Mik} = \min\{j \in C_{Mik}\}$. Let P_{M0} denote the set of continuous variables in the model. Sort model's parameters in the following order:

1. β_{00},
2. β_{j0}: $j \in P_{M0} \setminus \{0\}$,
3. $\beta_{j_{Mik}k}$ for $i = 1, \ldots, i_k$, $i \neq 1$, $k \in N \setminus \{0\}$,
4. β_{j0}: $j \in N_0 \setminus P_{M0}$,
5. β_{jk}, $j \in C_{Mik} \setminus \{j_{Mik}\}$, $k \in N \setminus \{0\}$.

Sort columns of model matrix $\mathbf{X}$ in the same way as vector $\boldsymbol{\beta}$.

Example 1. As an illustrative example consider a full model $F = (P_{F0}, P_{F1}, P_{F2})$, where

$$P_{F0} = \{1, 2\},\ P_{F1} = (\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}, \{7\}, \{8\}),$$
$$P_{F2} = (\{1\}, \{2\}, \{3\})$$

and $p_0 = 2, p_1 = 8, p_2 = 3, p = 12$. We denote a feasible model with 7 elementary constraints: $\beta_{10} = 0,\ \beta_{21} = 0,\ \beta_{71} = 0,\ \beta_{31} = \beta_{51},\ \beta_{41} = \beta_{61},\ \beta_{41} = \beta_{81},\ \beta_{22} = 0$ as $M = (P_{M0}, P_{M1}, P_{M2})$, where:

$$P_{M0} = \{2\},\ P_{M1} = (\{1, 2, 7\}, \{3, 5\}, \{4, 6, 8\}),\ P_{M2} = (\{1, 2\}, \{3\}).$$

Constraint matrix in regular form for model M, where each row corresponds to one of the 7 elementary constraints, is:

$$\mathbf{A}_{0M} = \begin{array}{c} \begin{array}{cccccccccccc} \beta_{00} & \beta_{20} & \beta_{31} & \beta_{41} & \beta_{32} & \beta_{10} & \beta_{21} & \beta_{71} & \beta_{51} & \beta_{61} & \beta_{81} & \beta_{22} \end{array} \\ \left[\begin{array}{cccccccccccc} 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right] \end{array}.$$

and after inverting matrix $\mathbf{A}_M^{-1}$ is obtained

$$\mathbf{A}_M^{-1} = \left[\ \mathbf{A}_M^1 \ \middle|\ \mathbf{A}_M^0\ \right]$$

$$= \begin{array}{ccccc|ccccccc}
\beta_{00} & \beta_{20} & \beta_{31} & \beta_{41} & \beta_{32} & \beta_{10} & \beta_{21} & \beta_{71} & \beta_{51} & \beta_{61} & \beta_{81} & \beta_{22} \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{array}$$

Notice that for regular constraint matrix $\mathbf{Z}_M$ is the full model matrix $\mathbf{X}$ with appropriate columns deleted or added to each other.

Appendix B: Detailed description of step 3 of the DMR algorithm

Since step 3 of the DMR algorithm needs complicated notations concerning hierarchical clustering, we decided to present them in the Appendix for the interested reader. In particular, we show here how the cutting heights vector $\mathbf{h}$ and matrix of constraints $\mathbf{A_0}$ are built.

Let us define vectors $\mathbf{a}(1, j, k)$ and $\mathbf{a}(i, j, k)$ (corresponding to the elementary constraints, being building blocks for $\mathbf{A_0}$) such that:

$$\mathbf{a}(1, j, k) = [a_{st}(j, k)]_{\substack{s \in N_t \\ t \in N}},\ a_{st}(j, k) = \mathbb{1}(s = j, t = k), \tag{16}$$

$$\mathbf{a}(i, j, k) = [a_{st}(i, j, k)]_{\substack{s \in N_t \\ t \in N}},\ a_{st}(i, j, k) = \mathbb{1}(s = i, t = k) - \mathbb{1}(s = j, t = k). \tag{17}$$

For each step s of the hierarchical clustering algorithm we use the following notation for the partitions of set $\{1\} \cup N_k = \{1, 2, \dots, p_k\}$:

$$P_{sk} = \{C_{isk}\}_{i=1}^{p_k - s + 1},\ s = 1, \dots, p_k.$$

We assume complete linkage clustering:

$$\begin{aligned} & d\left(C_{i_{s+1}, s+1, k} = C_{i_s sk} \cup C_{j_s sk}, C_{j_{s+1}, s+1, k} = C_{o_s sk}\right) \\ & = \max\left\{d\left(C_{i_s sk}, C_{o_s sk}\right), d\left(C_{j_s sk}, C_{o_s sk}\right)\right\}. \end{aligned}$$

Cutting heights in steps $s = 1, \dots, p_k - 1$ are defined as:

$$h_{sk} = \min_{i \neq j} d\left(C_{isk}, C_{jsk}\right).$$

Let us denote vector $\tilde{\mathbf{a}}_{sk}$ as an elementary constraint corresponding to cutting height h_{sk}, where:

$$\begin{aligned}\tilde{\mathbf{a}}_{sk} &= \mathbf{a}(i_*, j_*, k),\ i_* = \min_{i \in C_{i_1 sk}} i,\ j_* = \min_{j \in C_{j_1 sk}} j \text{ and } (i_1, j_1)\\ &= \arg\min_{i \neq j} d\left(C_{isk}, C_{jsk}\right).\end{aligned}$$

Step 3 of the algorithm can be now rewritten:

Combine vectors of cutting heights: $\mathbf{h} = [0, \mathbf{h}_0^T, \mathbf{h}_1^T, \ldots, \mathbf{h}_l^T]^T$, where $\mathbf{h}_0$ is vector of cutting heights for constraints concerning continuous variables and 0 corresponds to model without constraints:

$$\mathbf{h}_k = [h_{sk}]_{s=1}^{p_k - 1},\ k \in N \setminus \{0\} \text{ and } \mathbf{h}_0 = [0, t_{110}^2, t_{120}^2, \ldots, t_{1p_0 0}^2]^T.$$

Sort elements of $\mathbf{h}$ in increasing order getting $\mathbf{h}_{:} = [h_{m:p}]_{m=1}^{p}$ and construct $(p-1) \times p$ matrix of constraints

$$\mathbf{A}_0 = [\tilde{\mathbf{a}}_{2:p}, \tilde{\mathbf{a}}_{3:p}, \ldots, \tilde{\mathbf{a}}_{p:p}]^T,$$

where $\tilde{\mathbf{a}}_{m:p}$ is the elementary constraint corresponding to cutting height $h_{m:p}$. Then proceed as described in Algorithm 1.

Appendix C: Recursive formula for RSS in a nested family of linear models

In this section we show some implementation facts concerning the DMR algorithm. In particular an effective way of calculation of residual sums of squares for nested models using QR decompositions is discussed.

Let us consider a linear model with linear constraints:

$$\mathcal{L} = \{\boldsymbol{\beta} \in \mathbb{R}^p, \mathbf{A}_0 \boldsymbol{\beta} = \mathbf{0}\}, \tag{18}$$

where $\mathbf{A}_0$ is $(p-q) \times p$ constraint matrix. The objective is to calculate residual sum of squares $RSS = \|\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\|^2$. QR decomposition of the model matrix is performed

$$\mathbf{X} = \mathbf{QR},$$

where $\mathbf{Q}$ is $n \times p$ orthogonal matrix and $\mathbf{R}$ is $p \times p$ upper triangular matrix. Let us denote $\mathbf{S} = \mathbf{R}^{-T} \mathbf{A}_0^T$, then

$$\mathbf{Q}^T \mathbf{y} = \mathbf{R}\boldsymbol{\beta}^* + \mathbf{Q}^T \boldsymbol{\varepsilon} \text{ and } \mathbf{S}^T \mathbf{R} \boldsymbol{\beta}^* = \mathbf{0}.$$

After substitution $\mathbf{z} = \mathbf{Q}^T \mathbf{y}$, $\boldsymbol{\gamma}^* = \mathbf{R}\boldsymbol{\beta}^*$, $\boldsymbol{\eta} = \mathbf{Q}^T \boldsymbol{\varepsilon}$ we get

$$\mathbf{z} = \boldsymbol{\gamma}^* + \boldsymbol{\eta} \text{ and } \mathbf{U}^T \mathbf{W}^T \boldsymbol{\gamma}^* = \mathbf{0}, \tag{19}$$

where $\mathbf{W}$ and $\mathbf{U}$ are respectively $p \times (p-q)$ orthogonal matrix and $(p-q) \times (p-q)$ upper triangular matrix from the QR decomposition of matrix $\mathbf{S}$. We have

$$\mathbf{W}^T \boldsymbol{\gamma}^* = \mathbf{U}\mathbf{U}^T \mathbf{W}^T \boldsymbol{\gamma}^* = \mathbf{0}.$$

Let us denote $\overline{\mathbf{W}}$ as orthogonal complement of $\mathbf{W}$ to matrix with dimensions $p \times p$. We multiply equation (19) by $[\overline{\mathbf{W}}, \mathbf{W}]$:

$$[\overline{\mathbf{W}}, \mathbf{W}]^T \mathbf{z} = [\overline{\mathbf{W}}, \mathbf{W}]^T \boldsymbol{\gamma}^* + [\overline{\mathbf{W}}, \mathbf{W}]^T \boldsymbol{\eta} \text{ and } \mathbf{W}^T \boldsymbol{\gamma}^* = 0.$$

Therefore the OLS estimator $\widehat{\gamma}$ of $\boldsymbol{\gamma}^*$ with constraints satisfies the following equation

$$\begin{bmatrix} \overline{\mathbf{W}}^T \mathbf{z} \\ 0 \end{bmatrix} = [\overline{\mathbf{W}}, \mathbf{W}]^T \widehat{\gamma}. \tag{20}$$

Multiplying (20) by $[\overline{\mathbf{W}}, \mathbf{W}]$, we obtain $\overline{\mathbf{W}}\,\overline{\mathbf{W}}^T \mathbf{z} = \widehat{\gamma}$, then

$$(\mathbb{I} - \mathbf{W}\mathbf{W}^T)\mathbf{z} = \widehat{\gamma} = \mathbf{R}\widehat{\boldsymbol{\beta}}.$$

Let $\overline{\mathbf{Q}}$ be an orthogonal complement of $\mathbf{Q}$ to matrix with dimensions $n \times n$. The residual sum of squares for the model with linear constraints (18) can now be written as

$$\begin{aligned} RSS_M &= \|\overline{\mathbf{Q}}^T y\|^2 + \|\mathbf{Q}^T(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_M)\|^2 = \|\mathbf{y}\|^2 - \|\mathbf{z}\|^2 + \|\mathbf{Q}^T\mathbf{y} - \mathbf{R}\widehat{\boldsymbol{\beta}}_M\|^2 \\ &= \|\mathbf{y}\|^2 - \|\mathbf{z}\|^2 + \|\mathbf{W}\mathbf{W}^T\mathbf{z}\|^2 = \|\mathbf{y}\|^2 - \|\mathbf{z}\|^2 + \|\mathbf{W}^T\mathbf{z}\|^2 \\ &= \|\mathbf{y}\|^2 - \|\mathbf{z}\|^2 + \sum_{m=1}^{p-q} (\mathbf{w}_m^T \mathbf{z})^2, \end{aligned} \tag{21}$$

where $\mathbf{w}_m$ is the m-th column of $\mathbf{W}$.

Denote by $(\mathbf{A}_0)_{m,p}, \mathbf{S}_{m,p}, \mathbf{W}_{m,p}$ and $\mathbf{U}_{m,p}$ submatrices of $\mathbf{A}_0, \mathbf{S}, \mathbf{W}$ and $\mathbf{U}$ respectively, obtained by retaining first m rows and p columns. Let us consider a nested family of feasible models M_m, $m = 0, \ldots, p - q$ defined as

$$\mathcal{L}_{M_m} = \left\{\boldsymbol{\beta} \in \mathbb{R}^p, (\mathbf{A}_0)_{m,p}\boldsymbol{\beta} = \mathbf{0}\right\}.$$

For $m = 0, \ldots, p - q$ we have

$$\mathbf{S}_{p,m} = \mathbf{W}_{p,m}\mathbf{U}_{m,m},$$

because matrix $\mathbf{U}_{m,m}$ is upper triangular. Since $\mathbf{W}_{p,m}^T \mathbf{W}_{p,m} = \mathbb{I}$, then $\mathbf{W}_{p,m}\mathbf{U}_{m,m}$ is QR decomposition of $\mathbf{S}_{p,m}$. Then from equation (21) we get a recursive formula for residual sum of squares for nested models:

$$\begin{aligned} RSS_{M_0} &= \|\mathbf{y}\|^2 - \|\mathbf{z}\|^2, \\ RSS_{M_m} &= RSS_{M_{m-1}} + (\mathbf{w}_m^T\mathbf{z})^2 \text{ for } m = 1, \ldots, p-1. \end{aligned} \tag{22}$$

Appendix D: Proof of Theorem 1

D.1. Properties of orthogonal projection matrices

For a feasible model M let us define a following orthogonal projection matrix:

$$\overline{\mathbf{H}}_M = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}_{0M}^T \left(\mathbf{A}_{0M}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}_{0M}^T\right)^{-1} \mathbf{A}_{0M}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T.$$

Lemma 1. *We have*

$$\overline{\mathbf{H}}_M = \mathbf{H}_F - \mathbf{H}_M.$$

Proof. For simplicity of notations in the remainder of this subsection we omit subscript M. Let $\mathbf{Z}_1 = \mathbf{X}\mathbf{A}^1$, $\mathbf{Z} = \mathbf{X}\mathbf{A}^{-1}$ and $\mathbf{Z}_0 = \mathbf{X}\mathbf{A}^0$. We denote

$$\mathbf{G} = \begin{bmatrix} \mathbf{G}_{11} & \mathbf{G}_{10} \\ \mathbf{G}_{01} & \mathbf{G}_{00} \end{bmatrix} = \begin{bmatrix} \mathbf{Z}_1^T\mathbf{Z}_1 & \mathbf{Z}_1^T\mathbf{Z}_0 \\ \mathbf{Z}_0^T\mathbf{Z}_1 & \mathbf{Z}_0^T\mathbf{Z}_0 \end{bmatrix} = \mathbf{Z}^T\mathbf{Z} \text{ and } \mathbf{G}^{-1} = \begin{bmatrix} \mathbf{G}^{11} & \mathbf{G}^{10} \\ \mathbf{G}^{01} & \mathbf{G}^{00} \end{bmatrix}.$$

Note that

$$\mathbf{H}_F = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \mathbf{X}\mathbf{A}^{-1}(\mathbf{A}^{-T}\mathbf{X}^T\mathbf{X}\mathbf{A}^{-1})^{-1}\mathbf{A}^{-T}\mathbf{X}^T = \mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T.$$

Moreover

$$\begin{aligned}(\mathbf{A}_0(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}_0^T)^{-1} &= \left(\mathbf{A}_0\mathbf{A}^{-1}\left(\mathbf{A}^{-T}\mathbf{X}^T\mathbf{X}\mathbf{A}^{-1}\right)^{-1}\mathbf{A}^{-T}\mathbf{A}_0^T\right)^{-1} \\ &= \left[\begin{bmatrix} \mathbf{0} & \mathbb{I} \end{bmatrix}(\mathbf{Z}^T\mathbf{Z})^{-1}\begin{bmatrix} \mathbf{0} \\ \mathbb{I} \end{bmatrix}\right]^{-1} = (\mathbf{G}^{00})^{-1}\end{aligned}$$

and

$$\mathbf{A}_0(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \mathbf{A}_0\mathbf{A}^{-1}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{A}^{-T}\mathbf{X}^T = \mathbf{A}_0\mathbf{A}^{-1}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T.$$

Then we get from the Schur complement:

$$\begin{aligned}\mathbf{H}_F - \mathbf{H}_M &= \mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T - \mathbf{Z}_1(\mathbf{Z}_1^T\mathbf{Z}_1)^{-1}\mathbf{Z}_1^T = \mathbf{Z}\mathbf{G}^{-1}\mathbf{Z}^T - \mathbf{Z}_1\mathbf{G}_{11}^{-1}\mathbf{Z}_1^T \\ &= \mathbf{Z}\mathbf{G}^{-1}\mathbf{Z}^T - \mathbf{Z}_1(\mathbf{G}^{11} - \mathbf{G}^{10}(\mathbf{G}^{00})^{-1}\mathbf{G}^{10})\mathbf{Z}_1^T \\ &= \begin{bmatrix} \mathbf{Z}_1 & \mathbf{Z}_0 \end{bmatrix}\begin{bmatrix} \mathbf{G}^{11} & \mathbf{G}^{10} \\ \mathbf{G}^{01} & \mathbf{G}^{00} \end{bmatrix}\begin{bmatrix} \mathbf{Z}_1^T \\ \mathbf{Z}_0^T \end{bmatrix} \\ &\quad - \begin{bmatrix} \mathbf{Z}_1 & \mathbf{Z}_0 \end{bmatrix}\begin{bmatrix} \mathbf{G}^{11} - \mathbf{G}^{10}(\mathbf{G}^{00})^{-1}\mathbf{G}^{10} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}\begin{bmatrix} \mathbf{Z}_1^T \\ \mathbf{0}^T \end{bmatrix} \\ &= \mathbf{Z}\begin{bmatrix} \mathbf{G}^{10} \\ \mathbf{G}^{00} \end{bmatrix}(\mathbf{G}^{00})^{-1}\begin{bmatrix} \mathbf{G}^{01} & \mathbf{G}^{00} \end{bmatrix}\mathbf{Z}^T \\ &= \mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\begin{bmatrix} \mathbf{0} \\ \mathbb{I} \end{bmatrix}(\mathbf{G}^{00})^{-1}\begin{bmatrix} \mathbf{0} & \mathbb{I} \end{bmatrix}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T \\ &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}_M^T\left(\mathbf{A}_M(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}_M^T\right)^{-1}\mathbf{A}_M(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \overline{\mathbf{H}}_M.\end{aligned}$$

□

D.2. Asymptotics for residual sums of squares

Lemmas concerning dependencies between residual sums of squares have similar construction to those described in Chen and Chen (2008). Let us introduce some simplifying notations. For two sequences of random variables U_n and V_n we write that $U_n <_P V_n$ if $\lim_{n\to\infty} \mathbb{P}\left(U_n < V_n\right) = 1$.

Residual sum of squares for model M can be decomposed into three parts

$$\begin{aligned} RSS_M &= \|\mathbf{y} - \mathbf{H}_M \mathbf{y}\|^2 = (\mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon})^T (\mathbb{I} - \mathbf{H}_M)(\mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}) \\ &= \boldsymbol{\beta}^{*T}\mathbf{X}^T(\mathbb{I} - \mathbf{H}_M)\mathbf{X}\boldsymbol{\beta}^* + 2\boldsymbol{\beta}^{*T}\mathbf{X}^T(\mathbb{I} - \mathbf{H}_M)\boldsymbol{\varepsilon} + \boldsymbol{\varepsilon}^T(\mathbb{I} - \mathbf{H}_M)\boldsymbol{\varepsilon}. \end{aligned}$$

When $T \subseteq M$ we have $\mathbf{H}_M\mathbf{X}\boldsymbol{\beta}^* = \mathbf{X}\boldsymbol{\beta}^*$ and $RSS_M = \boldsymbol{\varepsilon}^T(\mathbb{I} - \mathbf{H}_M)\boldsymbol{\varepsilon}$.

Lemma 2. *Assuming $p \prec n$ and $p \prec r_n$, we have*

$$\log \frac{RSS_T}{RSS_F} <_P \frac{r_n}{n}.$$

Proof. Observe that

$$\frac{RSS_T}{RSS_F} = 1 + \frac{RSS_T - RSS_F}{RSS_F} = 1 + \frac{p}{n}E_n,$$

where

$$E_n = \frac{\boldsymbol{\varepsilon}^T(\mathbf{H}_F - \mathbf{H}_T)\boldsymbol{\varepsilon}}{\boldsymbol{\varepsilon}^T(\mathbb{I} - \mathbf{H}_f)\boldsymbol{\varepsilon}} \cdot \frac{n}{p}.$$

Let us notice that $\mathbf{H}_F - \mathbf{H}_T$ is a matrix of an orthogonal projection with rank $p - |T|$. Therefore $W_1 = \boldsymbol{\varepsilon}^T(\mathbf{H}_F - \mathbf{H}_T)\boldsymbol{\varepsilon} \sim \sigma^2\chi^2_{p-|T|}$ and $W_2 = \boldsymbol{\varepsilon}^T(\mathbb{I} - \mathbf{H}_F)\boldsymbol{\varepsilon} \sim \sigma^2\chi^2_{n-p}$. Then we get

$$\mathbb{E}\left(\frac{W_1}{p}\right) = \frac{\sigma^2(p - |T|)}{p}, \ \mathrm{Var}\left(\frac{W_1}{p}\right) = \frac{2\sigma^4(p - |T|)}{p^2}$$

and since p grows monotonically with n we have either $p \xrightarrow{n\to\infty} \infty$, then $\mathrm{Var}\left(\frac{W_1}{p}\right) \xrightarrow{n\to\infty} 0$ and from Chebyshev's inequality $\frac{W_1}{p} \xrightarrow{n\to\infty} \sigma^2$ in probability or p is bounded, then $\frac{W_1}{p}$ is bounded in probability. Analogously for W_2 we have

$$\mathbb{E}\left(\frac{W_2}{n}\right) = \frac{\sigma^2(n - p)}{n}, \ \mathrm{Var}\left(\frac{W_2}{n}\right) = \frac{2\sigma^4(n - p)}{n^2}$$

and since $p \prec n$ from Chebyshev's inequality $\frac{W_2}{n} \xrightarrow{n\to\infty} \sigma^2$ in probability.

Therefore $E_n = O_P(1)$ and $\frac{RSS_T}{RSS_F} = 1 + O_P\left(\frac{p}{n}\right)$. Hence

$$\log\left(\frac{RSS_T}{RSS_F}\right) = \log\left(1 + \frac{p}{n}E_n\right) \leq \frac{p}{n}E_n = O_P\left(\frac{p}{n}\right) <_P \frac{r_n}{n}. \qquad \square$$

Lemma 3. *Assuming that $p \prec \Delta$ (Δ is defined in equation (13)) we have for all $\delta > 1$*

$$\min_{M \in \mathcal{M}_\mathcal{V}} \left(\log\left(\frac{RSS_M}{RSS_T}\right)\right) \geq_P \log\left(1 + \frac{\Delta}{\delta\sigma^2 \cdot n}\right).$$

Proof. Using the fact that

$$\frac{1}{n}RSS_T = \frac{\boldsymbol{\varepsilon}^T(\mathbb{I}-\mathbf{H}_T)\boldsymbol{\varepsilon}}{n} = \sigma^2 + o_P\big(1\big)$$

and denoting

$$RSS_M - RSS_T = \Delta_M + S_M + W_T - W_M,$$

where

$$\Delta_M = \boldsymbol{\beta}^{*T}\mathbf{X}^T(\mathbb{I}-\mathbf{H}_M)\mathbf{X}\boldsymbol{\beta}^*,\ S_M = 2\boldsymbol{\beta}^{*T}\mathbf{X}^T(\mathbb{I}-\mathbf{H}_M)\boldsymbol{\varepsilon},\ W_T = \boldsymbol{\varepsilon}^T\mathbf{H}_T\boldsymbol{\varepsilon}$$
$$\text{and } W_M = \boldsymbol{\varepsilon}^T\mathbf{H}_M\boldsymbol{\varepsilon}.$$

Note that

$$\Delta_M \geq \Delta,\ S_M \sim \mathcal{N}(0, 4\sigma^2\Delta_M),\ W_T \sim \sigma^2\chi^2_{|T|} \text{ and } W_M \sim \sigma^2\chi^2_{p-1}.$$

Using assumption, $\frac{S_M}{\Delta_M}$, $\frac{W_T}{\Delta_M}$ and $\frac{W_M}{\Delta_M}$ are $o_P\big(1\big)$ from Chebyshev's inequality. Since the dimension of the true model T is finite and independent of n, so is the number of models in $\mathcal{M}_\mathcal{V}$ and we have

$$\begin{aligned} RSS_M - RSS_T &= \Delta_M\left(1 + \frac{S_M}{\Delta_M} + \frac{W_T}{\Delta_M} - \frac{W_M}{\Delta_M}\right) \\ &= \Delta_M\Big(1 + o_P\big(1\big)\Big) \geq \Delta\Big(1 + o_P\big(1\big)\Big). \end{aligned}$$

As a result

$$\log\frac{RSS_M}{RSS_T} = \log\left(1 + \frac{RSS_M - RSS_T}{RSS_T}\right) >_P \log\left(1 + \frac{\Delta}{\delta\sigma^2 n}\right) \text{ for } \delta > 1. \quad \square$$

Lemma 4. *Assuming that $p \prec \Delta$ we have*

$$\max_{M\in\mathcal{M}_\mathcal{T}}\Big(\log RSS_M\Big) <_P \min_{M\in\mathcal{M}_\mathcal{V}}\Big(\log RSS_M\Big),$$

Proof. For $\delta > 1$ let us denote $a = \log\big(1 + \frac{\Delta}{\delta\sigma^2 n}\big)$, then from Lemma 3 we get

$$\begin{aligned} \min_{M\in\mathcal{M}_\mathcal{V}}\Big(\log RSS_M\Big) &>_P \log RSS_T + a \geq \max_{M\in\mathcal{M}_\mathcal{T}}\Big(\log RSS_M\Big) + a \\ &\geq \max_{M\in\mathcal{M}_\mathcal{T}}\Big(\log RSS_M\Big). \end{aligned} \quad \square$$

D.3. Ordering of squared t-statistics

In this section we show that ordering of models $M \in \mathcal{M}_\mathcal{T} \cup \mathcal{M}_\mathcal{V}$ with respect to squared t-statistics is equivalent to ordering them with respect to the values of residual sum of squares.

Let t_M, where $M \in \mathcal{M}_\mathcal{T} \cup \mathcal{M}_\mathcal{V}$ denote t-statistic for the full model with one elementary constraint $\mathbf{A}_{0M}\beta = 0$.

Lemma 5. *If $p \prec \Delta$, then*

$$\max_{M \in \mathcal{M}_{\mathcal{T}}} t_M^2 <_P \min_{M \in \mathcal{M}_{\mathcal{V}}} t_M^2.$$

Proof. From Lemma 1 we get that

$$RSS_M - RSS_F = \mathbf{y}^T(\mathbf{H}_F - \mathbf{H}_M)\mathbf{y} = \widehat{\boldsymbol{\beta}}^T \mathbf{A}_{0M}^T (\mathbf{A}_{0M}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}_{0M}^T)^{-1}\mathbf{A}_{0M}\widehat{\boldsymbol{\beta}},$$

where $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$. Hence for for each $M \in \mathcal{M}_{\mathcal{T}} \cup \mathcal{M}_{\mathcal{V}}$

$$\begin{aligned} t_M^2 &= \frac{(\mathbf{A}_{0M}\widehat{\boldsymbol{\beta}})^2}{\widehat{\mathrm{Var}}(\mathbf{A}_{0M}\widehat{\boldsymbol{\beta}})} = \frac{(\mathbf{A}_{0M}\widehat{\boldsymbol{\beta}})^2}{\mathbf{A}_{0M}\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\beta}})\mathbf{A}_{0M}^T} = \frac{(\mathbf{A}_{0M}\widehat{\boldsymbol{\beta}})^2}{\widehat{\sigma}^2\mathbf{A}_{0M}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}_{0M}^T} \\ &= \frac{RSS_M - RSS_F}{\widehat{\sigma}^2}, \end{aligned}$$

where $\widehat{\sigma}^2 = \frac{RSS_F}{n-|F|}$. Observe that $\mathbf{A}_{0M}$ is $1 \times |F|$ matrix, thus

$$t_M^2 = (n - |F|)\frac{RSS_M - RSS_F}{RSS_F},$$

and from Lemma 4 we get the conclusion. □

D.4. Correct ordering of constraints using hierarchical clustering

In this subsection we state conditions under which the true model T belongs to the path of nested models obtained in step 4 of the DMR algorithm.

Temporarily let us limit the analysis to a model consisting of one factor and no continuous variables. The true partition of set $\{1, \dots, p_1\}$ will be denoted by $P_1^* = (C_{i1}^*)_{i=1}^{|T|}$. We say that distance matrix $\mathbf{D} = [d_{ij}]_{ij}$ is consistent with the true partition if dissimilarity measures for elements within the same clusters are smaller than for elements from different clusters:

$$\max_{l \in \{1,\dots,|T|\}} \max_{i,j \in C_{l1}^*} d_{ij} = d^{true} < d^{false} = \min_{\substack{l_1,l_2 \in \{1,\dots,|T|\} \\ l_1 \neq l_2}} \min_{i \in C_{l_1 1}^*, j \in C_{l_2 1}^*} d_{ij}. \tag{23}$$

Let $P_{s1} = (C_{is1})_{i=1}^{p_1-s+1}$ denote a partition of set $\{1, \dots, p_1\}$ in step s of hierarchical clustering algorithm, $s = 1, \dots, p_1$. We will name aggregation of $C_{i_s s1}$ and $C_{j_s s1}$ in step s compatible with the true partition P_1^* if there exist $l \in \{1, \dots, |T|\}$, $i_{s+1} \in \{1, \dots, p_1 - s\}$ and $i_s \neq j_s$, $i_s, j_s \in \{1, \dots p_1 - s + 1\}$ such that

$$C_{i_{s+1}s+11} = C_{i_s s1} \cup C_{j_s s1}\ ,\ C_{i_{s+1}s+11} \subseteq C_{l1}^*.$$

Cutting height in step s is defined as $h_{s1} = d(C_{i_s s1}, C_{j_s s1})$ if $C_{i_s s1}$ and $C_{j_s s1}$ are aggregated in this step, $\mathbf{h}_1 = (h_{11}, \dots, h_{p_1-1,1})$.

Lemma 6. *Assuming that the linkage criterion of hierarchical clustering algorithm satisfies:*

$$\begin{aligned} & d\left(C_{i_{s+1}s+1k} = C_{i_s sk} \cup C_{j_s sk}, C_{j_{s+1}s+1k} = C_{o_s sk}\right) \\ & \quad = b \min\left\{d\left(C_{i_s sk}, C_{o_s sk}\right), d\left(C_{j_s sk}, C_{o_s sk}\right)\right\} \\ & \quad + (1-b) \max\left\{d\left(C_{i_s sk}, C_{o_s sk}\right), d\left(C_{j_s sk}, C_{o_s sk}\right)\right\}, \end{aligned} \tag{24}$$

where $b \in [0,1]$ and the dissimilarity matrix has property (23), then the cutting heights for aggregations compatible with P_1^* are lower than d^{true} and cutting heights for aggregations not compatible with P_1^* are larger than d^{false}.

Proof. From (23) if $|T| = p_1$ the statement holds trivially and if $|T| < p_1$ aggregation in the first step is compatible with P_1^*. We assume that in step s aggregation is compatible with the true partition with cutting height not greater than d^{true}. If aggregation of $C_{i_{s+1}s+1,1} = C_{i_s s1} \cup C_{j_s s1}$ and $C_{j_{s+1}s+1,1} = C_{o_s s1}$ is compatible with P_1^* then

$$h_{s1} = d\left(C_{i_{s+1}s+11}, C_{j_{s+1}s+11}\right) \leq \max\left(d\left(C_{i_s s1}, C_{o_s s1}\right), d\left(C_{j_s s1}, C_{o_s s1}\right)\right) \leq d^{true}.$$

If aggregation of $C_{i_{s+1}s+11} = C_{i_s s1} \cup C_{j_s s1}$ and $C_{j_{s+1}s+11} = C_{o_s s1}$ is not compatible with P_1^* then

$$h_{s1} = d\left(C_{i_{s+1}s+11}, C_{j_{s+1}s+11}\right) \geq \min\left(d\left(C_{i_s s1}, C_{o_s s1}\right), d\left(C_{j_s s1}, C_{o_s s1}\right)\right) \geq d^{false}.$$

Hence, cutting heights $h_{11}, \ldots, h_{p_1-|T|,1}$ not greater than d^{true} are used until all aggregations compatible with P_1^* are performed. We have $C_{p_1-|T|+11} = P_1^*$ and in steps $s = p_1 - |T| + 2, \ldots, p_1$ the true partition P_1^* is a subpartition of C_{s1} and cutting heights $h_{p_1-|T|+11}, \ldots, h_{p_1-11}$ are not less than d^{false}. □

Note that linkage criteria: single, complete and average satisfy assumption (24).

Proof of Theorem 1a. Let us denote the path of nested models from step 4 of the DMR algorithm by $J = \{M_0, \ldots, M_{p-1}\}$. The event of erroneous selection of the model by the DMR algorithm is a subset of a sum of three events:

$$\begin{aligned} \{\widehat{T} \neq T\} \subseteq & \{T \notin J\} \cup \{T \in J, \mathrm{GIC}_T \geq \min_{M \subsetneq T} \mathrm{GIC}_M\} \\ & \cup \{T \in J, \mathrm{GIC}_T \geq \min_{T \subsetneq M} \mathrm{GIC}_M\} \\ \subseteq & \{T \notin J\} \cup \{\mathrm{GIC}_T \geq \min_{M \subsetneq T} \mathrm{GIC}_M\} \cup \{\mathrm{GIC}_T \geq \min_{T \subsetneq M} \mathrm{GIC}_M\}. \end{aligned}$$

We will show that the probability of each of them tends to zero when $n \to \infty$.

Using Lemma 5 let us consider constant h_* such that

$$\max_{M \in \mathcal{M}_{\mathcal{T}}} t_M^2 <_P h_* <_P \min_{M \in \mathcal{M}_{\mathcal{V}}} t_M^2.$$

It is obvious that cutting heights for true constraints for continuous variables are smaller than h_* and for false ones greater than h_*. It also follows from

Lemma 5 that dissimilarity matrices used in the algorithm are consistent with the partitions for model T. Then, applying Lemma 6 for each factor, we get that the cutting heights for aggregations compatible with the true partitions are not greater than h_* and for incompatible ones not smaller than h_*. Hence, in the DMR algorithm accepting true constraints precede accepting false ones, for large n the probability that the true model lies on the path of nested models tends to 1.

Since $\min_{T \subsetneq M} RSS_M \geq RSS_F$ we have

$$\{\mathrm{GIC}_T \geq \min_{T \subsetneq M} \mathrm{GIC}_M\} \subseteq \{\log RSS_T \geq \log RSS_F + \frac{r_n}{n}\}$$

and from Lemma 2 we know that

$$\mathbb{P}\left(\log RSS_T \geq \log RSS_F + \frac{r_n}{n}\right) \xrightarrow{P} 0.$$

It is obvious that

$$\{\mathrm{GIC}_T \geq \min_{M \subsetneq T} \mathrm{GIC}_M\} \subseteq \{\log RSS_T \geq \min_{M \in \mathcal{M}_\mathcal{V}} \log RSS_M - \frac{|T| r_n}{n}\}.$$

Let us notice from assumptions of theorem that $\frac{|T|r_n}{n} \prec \frac{\Delta}{\delta\sigma^2 n + \Delta} \leq \log\left(1 + \frac{\Delta}{\delta\sigma^2 n}\right)$. Then

$$\left\{\frac{|T|r_n}{n} \geq \min_{M \in \mathcal{M}_\mathcal{V}} \log \frac{RSS_M}{RSS_T}\right\} \supseteq \left\{\log\left(1 + \frac{\Delta}{\delta\sigma^2 n}\right) \geq \min_{M \in \mathcal{M}_\mathcal{V}} \log \frac{RSS_M}{RSS_T}\right\}$$

and from Lemma 3 we know that

$$\mathbb{P}\left(\log\left(1 + \frac{\Delta}{\delta\sigma^2 n}\right) \geq \min_{M \in \mathcal{M}_\mathcal{V}} \log \frac{RSS_M}{RSS_T}\right) \xrightarrow{P} 0.$$

Hence, the DMR algorithm is a consistent model selection method. □

Proof of Theorem 1b. Let us denote

$$\mathbf{g}_n = \sqrt{n}\left(\widehat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^*\right) \text{ and } \mathbf{b}_n = \sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{\widehat{T}} - \boldsymbol{\beta}^*\right),$$

Notice that $\mathbf{g}_n = \mathbf{b}_n$ if $\widehat{T} = T$. From Theorem 1a

$$\mathbb{P}\left(\mathbb{1}(\widehat{T} \neq T) = 0\right) \xrightarrow{P} 1.$$

Since

$$\left\{\mathbb{1}(\widehat{T} \neq T) = 0\right\} \subseteq \left\{\mathbf{b}_n \mathbb{1}(\widehat{T} \neq T) = 0\right\},$$

hence $\mathbf{b}_n \mathbb{1}(\widehat{T} \neq T) \xrightarrow{P} 0$. From properties of the OLS estimator we have

$$\mathbf{g}_n \mathbb{1}(\widehat{T} = T) \xrightarrow{d} \mathbb{N}(0, \sigma^2 \mathbf{\Sigma}_T).$$

Henceforth, from multidimensional Slutsky's theorem we get

$$\mathbf{b}_n = \mathbf{b}_n \mathbb{1}(\widehat{T} \neq T) + \mathbf{b}_n \mathbb{1}(\widehat{T} = T) = \mathbf{b}_n \mathbb{1}(\widehat{T} \neq T) + \mathbf{g}_n \mathbb{1}(\widehat{T} = T) \xrightarrow{d} \mathbb{N}(0, \sigma^2 \mathbf{\Sigma}_T).$$

□

References

Bondell, H. D. and Reich, B. J. (2009). Simultaneous factor selection and collapsing levels in ANOVA. *Biometrics* **65** 169–177. MR2665858

Caliński, T. and Corsten, L. (1985). Clustering means in ANOVA by simultaneous testing. *Biometrics* 39–48.

Chen, J. and Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* **95** 759–771. MR2443189

Ciampi, A., Lechevallier, Y., Limas, M. C. and Marcos, A. G. (2008). Hierarchical clustering of subpopulations with a dissimilarity based on the likelihood ratio statistic: application to clustering massive data sets. *Pattern Analysis and Applications* **11** 199–220. MR2411393

Dayton, C. M. (2003). Information criteria for pairwise comparisons. *Psychological Methods* **8** 61–71.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics* **32** 407–499. MR2060166

Gertheiss, J. and Tutz, G. (2010). Sparse modeling of categorial explanatory variables. *Annals of Applied Statistics* **4** 2150–2180. MR2829951

Oelker, M.-R., Gertheiss, J. and Tutz, G. (2014). Regularization and model selection with categorical predictors and effect modifiers in generalized linear models. *Statistical Modelling* **14** 157–177. MR3192557

Pokarowski, P. and Mielniczuk, J. (2015). Combined l_1 and greedy l_0 penalized least squares for linear model selection. *Journal of Machine Learning Research* **16(5)**.

Porreca, R. and Ferrari-Trecate, G. (2010). Partitioning datasets based on equalities among parameters. *Automatica* **46** 460–465. MR2877094

Scott, A. and Knott, M. (1974). A cluster analysis method for grouping means in the analysis of variance. *Biometrics* 507–512.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 267–288. MR1379242

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2004). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67** 91–108. MR2136641

Tukey, J. W. (1949). Comparing individual means in the analysis of variance. *Biometrics* 99–114. MR0030734

Zhang, T. (2010). Analysis of multi-stage convex relaxation for sparse regularization. *Journal of Machine Learning Research* **11** 1081–1107. MR2629825

Zheng, X. and Loh, W.-Y. (1995). Consistent variable selection in linear models. *Journal of the American Statistical Association* **90** 151–156. MR1325122

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics* **36** 1509. MR2435443