Computational Statistics and Data Analysis 93 (2016) 421–435

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Robust groupwise least angle regression

Andreas Alfons [a,b,*], Christophe Croux [b], Sarah Gelper [c]

[a] Erasmus School of Economics, Erasmus Universiteit Rotterdam, PO Box 1738, 3000DR Rotterdam, The Netherlands
[b] ORSTAT Research Center, Faculty of Economics and Business, KU Leuven, Naamsestraat 69, 3000 Leuven, Belgium
[c] Rotterdam School of Management, Erasmus Universiteit Rotterdam, PO Box 1738, 3000DR Rotterdam, The Netherlands

ARTICLE INFO

Article history:
Received 30 April 2014
Received in revised form 9 February 2015
Accepted 9 February 2015
Available online 17 February 2015

Keywords:
Categorical variables
Model selection
Outliers
Time series

ABSTRACT

Many regression problems exhibit a natural grouping among predictor variables. Examples are groups of dummy variables representing categorical variables, or present and lagged values of time series data. Since model selection in such cases typically aims for selecting groups of variables rather than individual covariates, an extension of the popular least angle regression (LARS) procedure to groupwise variable selection is considered. Data sets occurring in applied statistics frequently contain outliers that do not follow the model or the majority of the data. Therefore a modification of the groupwise LARS algorithm is introduced that reduces the influence of outlying data points. Simulation studies and a real data example demonstrate the excellent performance of groupwise LARS and, when outliers are present, its robustification.

1. Introduction

In many applications of linear regression, there exists a natural grouping among the predictor variables. One common example is regression with categorical variables, where each categorical variable is represented by a group of dummy variables. Another example is regression with time series data, where typically not only the original series are considered in the model, but also several lags of each series. Furthermore, time series models frequently contain an autoregressive part, i.e., lags of the response are included as covariates. Such models are commonly referred to as autoregressive models with exogenous inputs, or ARX models for short. Note that in both situations, groups of covariates emerge from the measured variables.

With increasing availability of data sets containing a large number of variables, model selection continues to be a topic of high importance in regression analysis. Linear models that include a large set of variables tend towards having large variance, often resulting in poor prediction performance. Selecting only the important variables can therefore improve prediction accuracy. Furthermore, traditional regression methods cannot be applied if the number of variables is larger than the number of observations due to the rank deficiency of the design matrix.

Whenever the regression problem involves groups of covariates, variable selection methods should select these groups rather than individual covariates. This ensures that all information of a selected measured variable enters the model, which is in general not the case when selecting individual covariates (Yuan and Lin, 2006). In addition, retaining the groupwise structure in submodels allows for better interpretation of the results.

* Corresponding author at: Erasmus School of Economics, Erasmus Universiteit Rotterdam, PO Box 1738, 3000DR Rotterdam, The Netherlands.
E-mail addresses: alfons@ese.eur.nl (A. Alfons), christophe.croux@kuleuven.be (C. Croux), sgelper@rsm.nl (S. Gelper).

http://dx.doi.org/10.1016/j.csda.2015.02.007
0167-9473/

Concerning notation, let n denote the number of observations and p the total number of covariates from m predictor groups. Moreover, let $\boldsymbol{y} = (y_1, \ldots, y_n)'$ be the response variable, and $\boldsymbol{X}_j$ an $(n \times p_j)$ matrix corresponding to the jth predictor group, $j = 1, \ldots, m$, with $\sum_{j=1}^{m} p_j = p$. The regression problem with grouped predictor variables can then be written as

$$\boldsymbol{y} = \sum_{j=1}^{m} \boldsymbol{X}_j \boldsymbol{\beta}_j + \boldsymbol{\varepsilon}, \tag{1}$$

where $\boldsymbol{\beta}_j$ is a coefficient vector of size p_j, $j = 1, \ldots, m$, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$ are the random error terms. Our aim is to find a subset $J \subseteq \{1, \ldots, m\}$ of the important predictor groups such that only those are included in the regression model, which is equivalent to setting the coefficient vectors $\boldsymbol{\beta}_j$ with $j \notin J$ in (1) to zero vectors.

In the traditional variable selection setting with all $p_j = 1$, a considerable amount of research has been done. Popular methods are the least absolute shrinkage and selection operator (lasso; Tibshirani, 1996), least angle regression (LARS; Efron et al., 2004), and the nonnegative garrote (Breiman, 1995). All three methods have been adjusted by Yuan and Lin (2006) to handle grouped variables. Zhao et al. (2009) introduced a family of composite absolute penalty functions for grouped and hierarchical variable selection via penalized regression. Furthermore, a groupwise version of the lasso for logistic regression was developed by Meier et al. (2008). Breheny and Huang (2009) follow a different philosophy and introduced a penalized regression framework for bi-level variable selection with grouped variables, i.e., their method first selects the important groups of variables and then the important variables within those groups. Nevertheless, none of these contributions consider the problem of outlying data points. While many robust methods for model selection in the traditional setting are available (e.g. Ronchetti et al., 1997; Khan et al., 2007; McCann and Welsch, 2007; Salibian-Barrera and Van Aelst, 2008; Khan et al., 2010; Alfons et al., 2011, 2013), almost no work has been done on robust groupwise variable selection. Chen et al. (2010) apply a more robust version of the groupwise lasso based on a convex combination of L_1 and L_2 loss functions. However, their procedure is only robust against heavily-tailed errors, but not against leverage points, i.e., outliers in the predictor space.

This paper focuses on the LARS procedure, which produces a sequence of variables in the order of their predictive content. Khan et al. (2007) point out that only correlations are required for variable sequencing with LARS and propose robustified versions of LARS, referred to as RLARS. While those authors express LARS in terms of correlations, we propose to use an extension of LARS to grouped variables that is formulated in terms of R^2 measures from short regressions. Here the term short regressions refers to regressions that use only one of the predictor groups. As the groupwise LARS approach is sensitive to outliers, we propose a robustification of the procedure such that the influence of outliers is reduced. We focus on sequencing the groups of variables, i.e., obtaining a sequence of groups in the order of their importance that can be further investigated for model selection and estimation.

The rest of the paper is organized as follows. Groupwise LARS is discussed in Section 2. Its robustification is then introduced in Section 3. Simulation studies are performed in Section 4, and Section 5 contains a real data example. Finally, Section 6 concludes. Proofs and technical details on the algorithms can be found in the Appendix.

2. Groupwise least angle regression

First, we review the idea of least angle regression (LARS) in the traditional setting with non-grouped variables. It proceeds in the following stepwise fashion (for details on the LARS algorithm, see Efron et al., 2004):

First step. Find the predictor with the highest correlation to the response and add it to the set of active predictors.

$(k+1)$th step. Move along the equiangular direction among all active predictors until a new predictor has equal correlation with the current residual, and add that predictor to the active set. The key to the algorithm is that this step size can easily be computed.

We generalize LARS by reformulating it in terms of R^2 measures from short regressions. A short regression has only the variables belonging to one single group as covariates, hence it has a limited number of regressors. This is in contrast to the full regression, where all covariates are included. If p is large with respect to n, short regressions can be carried out, while the full regression may not. Our approach is similar to the groupwise LARS algorithm of Yuan and Lin (2006), but our algorithm allows for more groups to be sequenced. The key steps of the groupwise LARS algorithm are discussed in the following. A complete schematic overview of the algorithm including technical details is given in the Appendix. Let $\boldsymbol{z}_0$ be the standardized response and $\underline{\boldsymbol{X}}_j$, $j = 1, \ldots, m$, the groups of standardized covariates such that all variables have zero mean and unit variance. Furthermore, let $R^2(\boldsymbol{z} \sim \boldsymbol{X})$ denote the R^2 measure of least squares regression of the vector $\boldsymbol{z}$ on the variables given by the columns of the matrix $\boldsymbol{X}$, let A denote the active set, i.e., the index set of the sequenced predictor groups, and let the complement A^c denote the inactive set, i.e., the index set of the not yet sequenced predictor groups.

First step. Find the predictor group with the largest R^2 measure

$$R^2(\boldsymbol{z}_0 \sim \underline{\boldsymbol{X}}_j), \quad j = 1, \ldots, m, \tag{2}$$

and add its index to the active set A.