Computational Statistics and Data Analysis 75 (2014) 39–52

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Group subset selection for linear regression

Yi Guo [a,*], Mark Berman [a], Junbin Gao [b]

[a] CSIRO Computational Informatics, North Ryde, NSW 1670, Australia

[b] School of Computing and Mathematics, Charles Sturt University, Bathurst, NSW 2795, Australia

ARTICLE INFO

Article history:
Received 1 February 2013
Received in revised form 17 October 2013
Accepted 5 February 2014
Available online 11 February 2014

Keywords:
Subset selection
Group Lasso
Linear regression
Screening

ABSTRACT

Two fast group subset selection (GSS) algorithms for the linear regression model are proposed in this paper. GSS finds the best combinations of groups up to a specified size minimising the residual sum of squares. This imposes an l0 constraint on the regression coefficients in a group context. It is a combinatorial optimisation problem with *NP complexity*. To make the exhaustive search very efficient, the GSS algorithms are built on QR decomposition and branch-and-bound techniques. They are suitable for middle scale problems where finding the most accurate solution is essential. In the application motivating this research, it is natural to require that the coefficients of some of the variables within groups satisfy some constraints (e.g. non-negativity). Therefore the GSS algorithms (optionally) calculate the model coefficient estimates during the exhaustive search in order to screen combinations that do not meet the constraints. The faster of the two GSS algorithms is compared to an extension to the original group Lasso, called the constrained group Lasso (CGL), which is proposed to handle convex constraints and to remove orthogonality requirements on the variables within each group. CGL is a convex relaxation of the GSS problem and hence more straightforward to solve. Although CGL is inferior to GSS in terms of group selection accuracy, it is a fast approximation to GSS if the optimal regularisation parameter can be determined efficiently and, in some cases, it may serve as a screening procedure to reduce the number of groups.

1. Introduction

In some regression problems it is natural to fit "groups" of explanatory variables at the same time. For example, in a birth weight data study (Yuan and Lin, 2006, Section 8), the authors used third-order polynomials to account for the non-linear effects of both mother's age and weight. Another example (Jacob and Obozinski, 2009, Section 9.4) is pathway analysis using biologically meaningful gene groups instead of individual genes. In this paper, we focus on a mineral spectroscopy application which motivated our research.

Spectroscopy is a fast surrogate for chemistry. The shortwave infrared (SWIR) reflectance spectrum of a mixture of minerals can be well approximated (after suitable transformation) by a linear combination of the spectra of its pure components. The coefficients in the linear model represent (in a semi-quantitative sense) the relative amounts of the corresponding mineral, so it makes sense to constrain them to be non-negative. Any single spectrum of a mixture contains only an identifiably small subset of the minerals in the library. For instance, our SWIR library contains multiple samples of 60 materials (mostly minerals). However, we and our specialist geologist collaborators have never observed a spectroscopic mixture in the SWIR of more than four materials. *The most important objective in this problem is, not to be able to predict spectra, but to identify*

* Corresponding author.
E-mail addresses: yi.guo@csiro.au (Y. Guo), mark.berman@csiro.au (M. Berman), jbgao@csu.edu.au (J. Gao).

http://dx.doi.org/10.1016/j.csda.2014.02.005
0167-9473/

correctly which minerals are in the mixture. A secondary objective is to estimate as accurately as possible the relative amounts of the minerals in the mixture.

Fig. 1 shows the 12 spectra in each of the Kaolinite_WX and Gypsum classes in our SWIR library after background removal and scaling. This will be further discussed in Section 5.1. Typically, we use the first principal component (PC) of the scaled and background removed spectra in a class to represent the whole class in a linear model. However, as can be seen in Fig. 1, the relative variation within the Gypsum class is much greater than that within the Kaolinite_WX class. This suggests that for classes such as Gypsum, one PC cannot adequately describe the whole class. Therefore it is useful to model some materials with multiple PC's to capture their greater variation. This results in the requirement of fitting several PC's together as a group to represent a material in the model. Furthermore, when the number of groups is large, it is often that only a small number of groups is needed to model the response variable, and so the question of how to choose the "best" groups arises.

We will consider group subset selection in the *linear* regression context. Formally, we assume that there are G groups of D-dimensional variables, $\mathbf{x}_{gi}$, $i = 1, \ldots, N_g$, $g = 1, \ldots, G$ (where N_g is the number of variables in group g and $\sum_g N_g = N$), and a D-dimensional response variable $\mathbf{y}$. Further, there is a subset of the G groups of (small) size M ($M \leq G$), $\mathcal{A}_M \in \{1, \ldots, G\}$, whose linear combination is a good model for $\mathbf{y}$. This can be represented as the linear model

$$\mathbf{y} = \sum_{g \in \mathcal{A}_M} \sum_{i=1}^{N_g} \beta_{gi} \mathbf{x}_{gi} + \boldsymbol{\epsilon}, \tag{1.1}$$

where $\boldsymbol{\epsilon}$ represents the error. In regard to estimation of β_{gi}, the regression coefficient for variable $\mathbf{x}_{gi}$, we consider ordinary least squares (OLS) minimising the following residual sum of squares (RSS)

$$\text{RSS} = \left\| \mathbf{y} - \sum_{g \in \mathcal{A}_M} \sum_{i=1}^{N_g} \beta_{gi} \mathbf{x}_{gi} \right\|_2^2. \tag{1.2}$$

Group subset selection (GSS) for OLS is defined as choosing $\mathcal{A}_M$ and coefficients β_{gi} to minimise (1.2) for given M.

At various points in the paper, it will be convenient to consider the non-group version of (1.2), i.e. given M, choose the subset of M *variables* $\mathcal{A}_M$ and coefficients, β_i, $i \in \mathcal{A}_M$ to minimise

$$\text{RSS} = \left\| \mathbf{y} - \sum_{i \in \mathcal{A}_M} \beta_i \mathbf{x}_i \right\|_2^2. \tag{1.3}$$

We will henceforth call minimisation of (1.3) *variable* subset selection (VSS). If we collect all the regression coefficients in a vector $\boldsymbol{\beta}$, the condition of $|\mathcal{A}_M| = M$ is actually equivalent to $\|\boldsymbol{\beta}\|_{l0} = M$ where $\|\mathbf{x}\|_{l0}$ is the number of nonzero entries in the vector $\mathbf{x}$.

Both VSS and GSS are combinatorial optimisation problems and therefore *NP hard* to solve. Of course, another "complication" of GSS is how to choose M, i.e. model selection. This issue will be discussed briefly later in this paper. *However the main focus will be on efficient methods of minimising* (1.2), *given* M.

Approximation of (1.2) has been considered by Yuan and Lin (2006). Generalising the Lasso (Tibshirani, 1996), they develop the group Lasso (GL) as a convex relaxation of (1.2). This selects groups via the use of $l1/l2$ regularisation on regression coefficients replacing the stringent $l0$. GL transforms the difficult combinatorial optimisation problem to an easier convex programming at the cost of introducing a bias due to the regularisation. This may not be a serious problem in prediction problems. However, in some applications such as mineral spectroscopy, where the central problem is correct identification rather than prediction, the bias will often decrease classification accuracy. For example Guo and Berman (2012) demonstrated that, in the variable selection context, direct minimisation of (1.3) usually has lower error rates in the identification of mineral composition than the Lasso does.

Because GL inherits the regularisation bias of the Lasso, we decided to generalise VSS to GSS. There are three primary aspects to the generalisation: (i) the obvious extension of variables to groups, (ii) optimisation of the code, which makes it significantly faster in certain situations, and (iii) the optional inclusion of some constraints on some of the coefficients.

The paper is structured as follows. Section 2 consists of a brief literature review of subset selection algorithms in the non-group context. Two fast GSS algorithms optimising (1.2) directly are presented in Section 3. We compare GSS with the GL algorithm of Yuan and Lin (2006) which is the best known regularised group selection algorithm. However, GL has two shortcomings: (i) it requires variables within each group to be orthogonal and (ii) it does not allow optional convex constraints on the coefficients of some variables. Therefore, in Section 4, we present an extension of GL, called the constrained group Lasso (CGL), which overcomes these problems. GSS, VSS and CGL are compared in terms of accuracy and speed in Section 5, with the aid of the spectroscopic unmixing problem. The work is summarised and possible future research is discussed in Section 6.

2. Literature review

In the non-group (VSS) context, a number of algorithms have been proposed to minimise (1.3) either exactly or approximately. Miller's fast variable selection algorithm (Miller, 2002) finds the *exact* minimum efficiently by utilising QR decomposition and branch-and-bound techniques. It can solve moderately sized variable subset selection problems in reasonable time. For example, choosing the best 4 out of 100 variables can be solved in seconds on a modern personal computer. A