Another look at subset selection using linear least squares SPECIAL CHAPTERS OF ARTIFICIAL INTELLIGENCE - RESEARCH REVIEW

A PREPRINT

Allan J. Miller
Retired,
1 Creswick St.,
Brighton E.,
Vic. 3187, Australia

Madalina-Alina Racovita,
Computational Optimization Student,
Faculty of Computer Science,
"Alexandru Ioan Cuza"
University of Iasi

January 16, 2020

ABSTRACT

Using QR decomposition, subset selection can be pursued when encountering models that respect least squares principles. The author derives a new stopping rule for the previously mentioned algorithm. In the appendix of this research article an estimator for squared projections is presented.

***K*eywords** Subset selection · Linear regression · Stopping rules · Overfitting

1 Introduction

Classical stepwise regression algorithm proposed by Efroymson would have a large execution time when the cardinal of the independent predictors is highly increased. In this case, the number of models that have to be checked in the entire algorithm is very large. The presence of stopping rules is self-understood, but finding these filtering conditions represents a difficult problem due to the fact that they are strictly dependent on the variables chosen in the current submodel and on the variables that are predisposed to be selected in a new iteration. QR factorizations give a useful insight into the extent of biases introduced while searching for best-fitting models.

2 Orthogonal projections and their properties

In linear algebra, a **QR decomposition** [1], also known as a QR factorization is a decomposition of a matrix A into a product $A = QR$ of an orthogonal matrix Q (i.e. $Q^T = Q^{-1}$) and an upper triangular matrix R. There are multiple approaches for transforming the columns of a X matrix of dimension $n * k$ (n is the number of observations from the considered sample and k is the number of predictors) in k orthogonal variables:

$$Q^T X = \begin{pmatrix} R \\ 0 \end{pmatrix}, \tag{1}$$

where R is a square matrix of $k * k$ dimension, and Q is a $n * n$ matrix. **Principal components analysis** or **Singular Value Decomposition** are recurrent preferences, but in both cases each new orthogonal variable is a linear combination of all the original predictors. On balance, there are **algorithms of orthogonalization** for which the first orthogonal variable involves only one predictor from the initial set of independent variables, the second is a linear combination of two of them etc.

Among these algorithms there were mentioned the **Givens** algorithm that computes QR factorizations by using rotation matrices whose columns form an orthogonal basis. The algorithm can be used as well in finding the eigenvalues of a matrix; the **Gram-Schmidt** algorithm (or **Laplace** method) that computes the QR decomposition on columns and it is also an algorithm of orthonormalisation of a basis. Last but not least, the **Householder** algorithm which uses reflexion matrices. The QR factorizations are particularly suited for subset selection.

The following relations are inferred, considered $X = (X_1, X_2, ..., X_k)$, $Q = (Q_1, Q_2, ..., Q_3)$, $R = \{r_{ij}\}$:

$$X = Q\begin{pmatrix} R \\ 0 \end{pmatrix} \tag{2}$$

$\rightarrow X1 = r_{11}Q_1$
$X2 = r_{21}Q_1 + r_{22}Q_2$
$X3 = r_{31}Q_1 + r_{32}Q_2 + r_{33}Q_3$ **etc.**

From the above relations, we deduct the fact that each X_i is a linear combination of the first i orthogonal columns of the matrix Q. The following relation:

$$Q^T y = t \tag{3}$$

designates t consisting of the **least-squares projections of** y, on the orthogonal directions in consecutive columns of Q, or more clearly the least-squares estimates of the regression coefficients of the variable Y against the columns of Q. The **regression coefficients**, β, of Y against X are estimated by solving the following equation:

$$R\widehat{\beta} = t, \tag{4}$$

equation which derives from solving a system of equations (many times being overdetermined, i.e. the dataset has more observations than predictors) by using the QR decomposition:

$$X\beta = y \rightarrow QR\beta = y| \cdot_{right} Q^T \rightarrow I_n R\beta = Q^T y| \cdot_{right} R^{-1} \rightarrow \beta = R^{-1}Q^T y \tag{5}$$

2.1 Properties of the QR decomposition useful in subset selection

1. **The matrix** R **is the Cholesky factor of** $X^T X$, i.e. $X^T X = R^T R$. The Cholesky decomposition [2] is the LU decomposition for symmetric and positively defined matrices: $A = L^T L$.
2. **The first p rows and columns of** R **are the Cholesky factor for the first** p **rows and columns of** XX^T. This means that for the regression of Y against the first p columns of X, we can use the same calculations as for the full regression but just *'forget'* the last p rows and columns.
3. **The order of the variables in the Cholesky factor,** R**, can easily be changed** to derive regressions of y upon other subsets of X-variables.
4. If the data satisfies the relation $y = X\beta + \epsilon$, where $\epsilon \sim \mathcal{N}(\mu, \sigma^2)$, then $t \sim \mathcal{N}(\mu, \sigma^2)$. This is derived as a consequence of the **Central Limit Theorem**.
5. If only the p elements of β are non-zero, then **the expected values of the last** $(n - p)$ **elements of** t **is to be zero**, i.e.:

$$\beta = (\beta_1, \beta_2, ..., \beta_p, ...)\ where\ \beta_i \neq 0, \forall i = \overline{1,p} \rightarrow t_j = 0, \forall j = \overline{p+1,n}\ where\ t = (..., t_{p+1}, ..., t_n) \tag{6}$$

6. **The square of the** i**-th projection,** t_i**, is the reduction of the residual sum of squared** when the variable in the position i is added to the linear model containing the first $(i - 1)$ variables.
7. **Rearranging the order of the variables**, for instance between positions i and j, will change all elements in those rows, and interchange elements in the same columns.

The author illustrates all of these properties by using the **STEAM** dataset from *Draper&Smith*, when predicting the values of a dependent variable y, which represents the number of pounds of steam used in each of 25 consecutive months.

3 Stopping rules

This chapter of the article presents the idea in which if the sample size is increased, i.e. the value of n is increased, the elements values of R, which is a $k \cdot k$ matrix and the size of t which is a vector of n elements are increased approximately with $\sqrt{n}$. More precisely, if there are selected independently and with the same probability samples from a population that respects a multivariate distribution of finite variance, the elements $\{r_{ij}|i - fixed, j = \overline{1,k}\}$ and the elements of t will be proportional with $\sqrt{n+1-i}$. This value $\sqrt{n+1-i}$ is proportional with the square root of the number of degrees of freedom.

In the article is considered a situation in which two or more predictors are highly correlated to each other, for example $X_4 = \textit{wind velocity}$ and $X_9 = \textit{the square of wind velocity}$, both variables extracted from steam dataset. If the variable X_4 appears after X_9 as columns in R matrix, then in the t vector,it will be assigned large projections for both of the variables. Because the expected value of the t_i increases as $\sqrt{n+1-i}$, the author propose **a solution for making the inclusion of the variables in the model gradually tougher as the sample size increases by proposing a new stopping rule.**

Let the predicted value be noted with $\widehat{y}$; it represents an elements from the vector $\widehat{Y}$, the vector with the regression predictions, computed after finding the coefficient values. By being given one observation from X, noted with x, then:

$$\widehat{y} = x^T \widehat{\beta} = x^T R^{-1} t = z^T t, \tag{7}$$

where $z^T R = x^T$. If for the current iteration it has been chosen a subset of p values, we order the independent variables, so that on the first positions in t are the corresponding values for the first p predictors. The last $(k-p)$ values of t will be completed with zeros.

The notation for this ordering will be the following: $E(t_i) = \tau_i$. The following relations are deducted:

$$bias = \sum_{i=1}^{p} z_i \{E(t_i) - \tau_i | selection + \sum_{i=p+1}^{k} z_i \tau_i\} \tag{8}$$

$$var = \sum_{i=1}^{p} z_i^2 var(t_i) \simeq \sigma^2 \sum_{i=1}^{p} z_i^2 \tag{9}$$

Using **(8)** and **(9)** and neglecting the first term for the bias since in many case it's expected to be zero, it is derived the squared error of prediction, or shortly **SEP**, as:

$$\sigma^2 \sum_{i=1}^{p} z_i^2 + \sum_{i=p+1}^{k} z_i \tau_i \tag{10}$$

If more variables are allowed into the model, then the first term increases with $\sigma^2 z_{p+1}^2$. Since the τ_i^2 is unknown, in the appendix it is presented an estimator for squared projections. Applying the shrinkage factor to **10**, it is obtained the estimation for **SEP**:

$$\sigma^2 \sum_{i=1}^{p} z_i^2 + C(\sum_{i=p+1}^{k} z_i t_i)^2 \tag{11}$$

A new stopping rule is derived by minimizing the above estimation with respect to the number of variables, p, in the subset.

4 Appendix. An estimator for squared projections

The appendix of the article presents the idea of estimating μ^2 (i.e. a **squared projection**), given a random variable x which is normally distributed, $x \sim N(\mu = ?, \sigma^2)$. The value of the variance is known, but the mean, not. It is considered an estimator, Cx^2 where C is a **shrinkage coefficient**, real and sub unitary, $C \in \mathbb{R}, 0 \leq C \leq 1$. The value of the parameter C has to be found, so that a squared error loss is minimized. From the normal distribution it is derived the following relation:

$$E(Cx^2 - \mu^2) = C^2(\mu^4 + 6\mu^2\sigma^2 + 3\sigma^4) - 2C\mu^2(\mu^2 + \sigma^2) + \mu^4 \tag{12}$$

If the above relation is differentiated with respect to C, it is obtained:

$$\frac{\partial E(Cx^2 - \mu^2)}{\partial C} = 2C(\mu^4 + 6\mu^2\sigma^2 + 3\sigma^4) - 2\mu^2(\mu^2 + \sigma^2), \tag{13}$$

relation that is minimized, if it is equaled to 0 and divided with σ^4 when:

$$C = \frac{(\frac{\mu^2}{\sigma^2})^2 + \frac{\mu^2}{\sigma^2}}{(\frac{\mu^2}{\sigma^2})^2 + 6\frac{\mu^2}{\sigma^2} + 3} \tag{14}$$

Since the value of C has to be found and this formula is dependent of the value of μ which is unknown, a substitution with the value of x is made. The following relation is derived:

$$\mu^2 = x^2 \frac{(\frac{x^2}{\sigma^2})^2 + \frac{x^2}{\sigma^2}}{(\frac{x^2}{\sigma^2})^2 + 6\frac{x^2}{\sigma^2} + 3} \tag{15}$$

The author emphasis the fact that for large x, $\mu^2 \simeq x^2 - 5\sigma^2$, and for decreased dimension of x, $\mu^2 \simeq \frac{x^4}{3\sigma^2}$

5 General thoughts based on the article. Personal opinion

The article was interesting in terms of numerical calculus and subset selection concepts. The manner in which it was written, was sometimes a bit confusing since for instance in the appendix there are not as many details as they should've been. The explanations for the new stopping rule are not that intuitive, from my personal point of view, since the lecture of the targeted chapter is a bit difficult to understand. I didn't get the idea form the first reading, by contrary, there were necessary 3 or 4 times to re-read that part so that I can finally understand it. All in all, it was an interesting paper, helping me to learn some useful properties regarding the usage of QR decompositions and subset selection.

References

[1] **QR factorization**: *https://en.wikipedia.org/wiki/QR_decomposition*

[2] **Cholesky Decomposition:** *https://en.wikipedia.org/wiki/Cholesky_decomposition*