Bioinformatics

Development of the R package 'L0adri' for best subset selection using iterative adaptive ridge procedure

Colaes Rob[1], Ding Mianyong[1] , Dubai Li[1], Medaer Margot[1], Prof. Dr. Wenseleers Tom[2]

[1]Department of Bioinformatics, KU Leuven, Leuven, Belgium, [2]Department of Biology, KU Leuven, Laboratory of Socioecology and Social Evolution, Leuven, Belgium

Abstract

Motivation: Lasso and best subset selection are variable selection methods for excluding irrelevant variables. These methods help to avoid overfitting of data and the presence of high variance when having high dimensional inference in generalized linear models. But these methods have the disadvantage of either producing biased coefficients or being non-convex and hard to solve. The newly developed R package "L0adri" will introduce the L_0-penalized generalized linear model for high dimensional inference. It makes use of a series of convex problems and implements the iteratively reweighted least squares.

Results: The newly designed package which implements the iterative adaptive ridge regression has been made more modular. It gives the option for box constraints and different solvers. It is applicable for high dimensional problems with solvers optimized for sparse or dense covariate design matrices.

Contact: tom.wenseleers@kuleuven.be

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The least squares method is often used for fitting a generalized linear model (GLM). But it can cause some difficulties in terms of the prediction accuracy and model interpretability (James *et al*., 2017). When the number of observations is lower than the number of variables, the model will overfit the data and cause high variance. A solution for this is thus constraining or shrinking the estimated coefficients of a model. To interpret a model correctly, it is useful to set estimated coefficients of variables that don't associated with the response variable to zero. That is why variable selection methods are introduced (James *et al*., 2017).

Alternative methods like subset selection and shrinkage methods for fitting a model to data are suggested. The subset selection method is a method that sets the irrelevant coefficient estimates to zero and uses the L_0-norm-penalized regression. The difficulty for the L_0-penalty is that it is an NP hard problem, and the optimization problem is difficult to solve due to its non-convex nature (Frommlet & Nuel, 2016). Another method that can be used to fit the model with high dimensionality is a shrinkage method like Lasso. It uses the L1-penalty. This method will shrink their coefficient estimates towards zero with the help of a tuning parameter. But the disadvantage is that it will produce biased coefficients.

The objective of this research is to develop an R package with a variable selection method that will implement iterative adaptive ridge regression which approximates the L_0-penalty. The implementation should be in a way that it is modular and can be used for high dimensional data.

2 Methods

In our R-package, an iterative adaptive ridge procedure is used to implement a L_0-penalty on a generalized linear model for best subset selection. The package uses different steps to implement this on the GLM and this will be explained in this section.

2.1 Generalized Linear Model (GLM) and Iterative Reweighted Least Squares

Our package implements the L_0-penalty on a generalized linear model. The goal of a generalized linear model is to specify a relationship between a response variable and multiple covariates. The model is a 'relaxed' extension to the linear regression model allowing for more general distributions for the response and allowing for a description of the variance instead of assuming constant variance for each covariate.

The GLM consists of two components: a random component and a systematic component. The random component determines the appropriate probability distribution for the response. It assumes it comes from a family of distribution, the exponential dispersion model (EDM). The EDM makes it possible for the GLM to fit a variety data types. We have:

$$y_i \sim EDM(\mu_i, \frac{\phi}{w_i})$$

The systematic component specifies the relationship between the explanatory variables and the mean of the response. The linear predictor (η) is linked to the mean by a monotonic link function g():

$$\eta = \beta_i + \sum \beta_j x_j$$
$$g(y) = \eta$$

The full GLM then becomes:

$$y_i \sim EDM(\mu_i, \frac{\phi}{w_i})$$
$$g(\mu_i) = o_i + \beta_0 + \sum \beta_j x_{ji}$$

(Hardin *et al.*, 2018) & (Dunn *et al.,* 2018)

Furthermore, the package uses Iterative Reweighted Least Squares (IRLS) to get a maximum likelihood estimation of the coefficients (β). The IRLS method iteratively solves a weighted least squares for a GLM. The method iteratively reweights a weighted least square solution to let it converge to an optimal solution. The weighted least squares minimize the least squares solution but applies weights on certain components. Solving this weighted least squares is the main step and the outcome is used in the next iteration.

Specifically, the weights are on the diagonal of a weights matrix W, which is in the first iteration equal to an identity matrix. From these weights and coefficients matrix A, a solution is found for X. From the solution a new weighted least squares error is determined, which then is used to get an updated weights matrix. From the new weights matrix, a new iteration can begin. This process is repeated until convergence (Sidney Burrus, 2012).

2.2 Iterative Adaptive Ridge Procedure

In the iterations of IRLS the coefficient estimates are penalized, and different methods for this exists. In general, the goal is to minimize the contrast C(β) penalized with a certain penalty and certain tuning parameter λ:

$$\widehat{\beta} \triangleq argmin_\beta [C(\beta) + \lambda \|\beta\|_L^q]$$

Multiple types for the contrast are possible, for example the residual sum of squares. Notice that depending on the value of q a different penalty and thus a different norm is applied. If q=1, the L_1-penalty is applied, which corresponds to Lasso-regression. If q=2 the L_2-penalty is applied, which corresponds to Ridge-regression, and if q=0 the L_0-penalty is applied. The advantage of using a L_0-penalty is that it is the strictest one, counts the number of non-zeroes and can perform subset-selection. However, computing this penalty is very challenging due to its non-convex nature and discontinuity, which makes it a NP-hard problem to solve. The L_1-norm and L_2-norm are less challenging to compute, explaining why it is often preferred over the L_0-penalty.

Thus, to solve this our package uses iteratively adaptive ridge procedure to approximate the penalty. In every iteration the procedure performs weighted ridge regression. The weights of the procedure are specifically constructed so that the imposed penalty converges towards the L_0-penalty. The weights will apply a larger penalty on coefficients that are near zero. The Adaptive Ridge procedure is defined as:

$$\beta^{(k)} \triangleq \arg\min_\beta F_{\lambda, w^{(k-1)}}(\beta)$$

$$F_{\lambda,w}(\beta) \triangleq C(\beta) + \frac{\lambda}{2}\beta^T diag(w)\beta = C(\beta) + \frac{\lambda}{2}\sum w_j \beta_j^2$$
$$w^{(k)} = \left(\left(\beta^{(k)}\right)^k + \delta^y\right)^{(q-2)/y}$$

(Frommlet & Nuel, 2016)

2.3 Solvers

To solve this implementation of the Adaptive Ridge procedure, our package contains two possible solver algorithms, a linear solver and a OSQP solver. The user can select which linear solver is to be used and whether box constraints should be applied. The goal of the solver is thus to find □ that minimizes the equation above.

The first solver is a linear solver, which is extended with clipping for implying box constraints, and contains two different methods. This method implements the Least Squares Conjugate Gradient method if the covariate matrix is sparse, and a linear solver method implemented by the solve function of the package base if the matrix is dense. The LSGD is excellent iterative method for large and sparse matrices to minimize a quadratic function (Stuetzle *et al.,* 2001).

The second option is the Operator Splitting solver for Quadratic Programs. This method reformulates the problem as a convex problem. It is very fast and robust algorithm that doesn't add certain constraints. The solver uses the Alternating Direction Method of Multipliers (ADMM) algorithm to solve minimization problems of the following form:

$$\frac{1}{2}x^T P x + q^T x$$
$$with: \; lower \leq Ax \leq upper$$

(Stellato *et al.*, 2018)

3 Results

3.1 Ground truth vs fitted value

The l0adri package was tested on the simulated spike data and the results are shown in figure 1. For all plots, the gray lines in the upper parts are simulated signals and the red vertical lines are the true coefficients of the simulated spikes. The yellow lines in the lower parts represent the fitted signal generated by the l0adri package and the blue vertical lines are the estimated coefficient. The great performance of the l0adri package can be seen from both figures, the fitted signals are quite similar to the simulated signals. And most of the estimated coefficients are close to the true coefficients. However, for some coefficients, there are small deviations of the position of peaks between the ground truth and estimated signals.

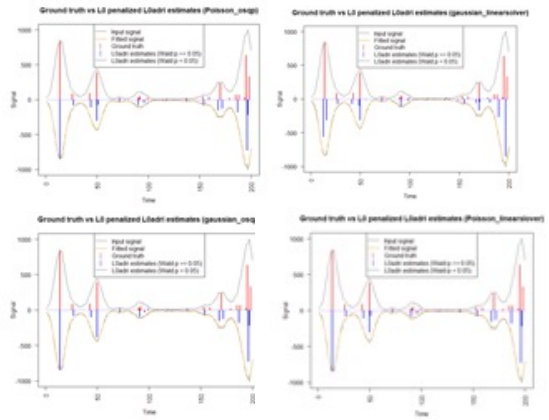

Figure 1. Ground truth vs fitted value (Upper left: Poisson distribution with OSQP solver, Upper right: gaussian distribution with linear solver, lower left: gaussian distribution with OSQP solver, lower right: Poisson distribution with linear solver).

3.2 Comparing with other packages

Next, we compared the performance of our package with other common packages with regularization methods. The glmnet package (Friedman, Hastie & Tibshirani, 2010) with the choice of lasso regression (L1 penalty) and l0learn (Hazime & Mazumder, 2020) with the penalty of L_0-L_2 were chosen firstly. The simulated data has 6 peaks (gray line), all the methods we used here only catch 5 peaks. As the estimated coefficients of l0adri with the choice of OSQP solver are the same as coefficients estimated by linear solver, the line of l0adri_OSQP is overlapped by the loadri_ls. For the three methods, they do not catch the first peak very well. For the second, fourth and fifth peaks, the estimates of three methods are quite close. And for the third and sixth peak, our package is much closer to the real coefficients of the peak.

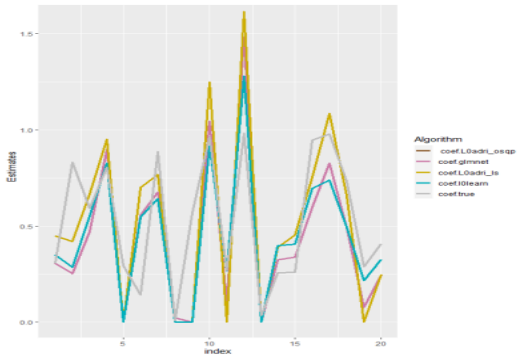

Figure 2. The comparison of estimated coefficients between l0learn, glmnet, l0adri.

We extended the comparison to more regularization packages oridinis package (Jared, 2018) and ncvreg package (Breheny & Huang J, 2011). Three choices of penalty of ordinis package were chosen which are minimax concave penalty (MCP), smoothly clipped absolute deviation (SCAD) and adaptive lasso (ALASSO). For the ncvreg, the MCP penalty was used. The accuracy, specificity and sensitivity of each package were estimated and compared (figure 3 and table 1).

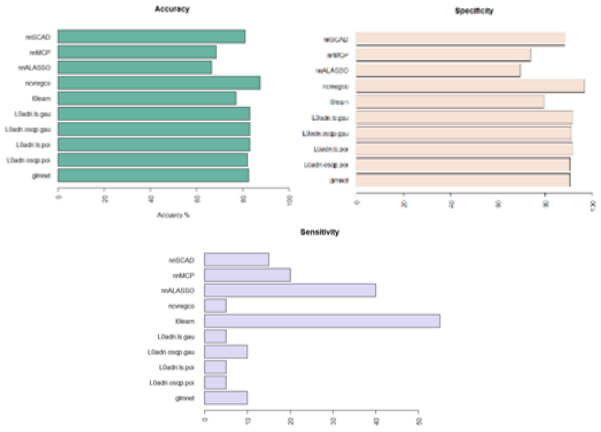

Figure 3. The accuracy, specificity and sensitivity of different packages. (nnSCAD: ordinis with SCAD penalty, nnMCP: ordinis with MCP penalty, nnALASSO: ordinis with ALASSO penalty, ncvregco:ncvreg package, l0learn: l0learn package,L0adri.ls.gau:l0adri package with gaussian distribution and linear solver, l0adri.OSQP.gau: l0adri package with gaussian distribution and OSQP solver, l0adri.ls.poi: l0adri package with Poisson distribution and linear solver, l0adri.OSQP.poi: l0adri with Poisson distribution and OSQP solver).

Table 1. The accuracy, specificity, and sensitivity of different packages.

methods	sensitivity %	accuracy %	specificity %'
glmnet	10	82.5	90.56
L0adri.osqp.poi	5	82	90.56
L0adri.ls.poi	5	83	91.67
L0adri.osqp.gau	10	83	91.11
L0adri.ls.gau	5	83	91.67
l0learn	55	77	79.44
ncvregco	5	87.5	96.67
nnALASSO	40	66.5	69.44
nnMCP	20	68.5	73.89
nnSCAD	15	81	88.33

Compared with other packages with regularization methods, our package shows high accuracy, high specificity and low sensitivity with the four combinations of parameters. The lower performance of sensitivity may be due to the criteria we used. We treated coefficients as negative if they are no larger than 0 and positive if the values of coefficients are larger than 0. The majority of the estimated coefficients are zero which are classified as negative. From figure 1, we can see the value of estimated coefficients are quite close to the real coefficient. However, there are some deviations of positions of peaks, leading to the low performance of sensitivity of our package. The l0learn shows the highest sensitivity, suggesting the L_0-L_2 penalty may be the possible extension to our package in the future.

Conclusion

Here we developed an R-package "L0adri" which performs best-subset selection using an iterative adaptive ridge procedure to approximate L_0-penalty. We based our work on the previous research group and introduced some extensions. There are multiple new features for the users to choose from. The automated cross-validation is built-in for tuning the regularization parameter lambda, and bootstrapping procedure for estimates. On top of the non-negativity constraints, the box constraints method was allowed for limiting the coefficients. Besides, the original implementations of different solvers were modularized, thus making it more readable and feasible to add more options. We test the performance of our package on

prediction the retention indices of compounds in the field of applied chemistry. It showed our package worked well compared to the other packages. But certainly, a better criterion needs to be proposed when measuring the deviance of prediction from the ground truth. This package is applicable for other high dimensional problems with solvers optimized for sparse or dense covariate design matrices.

Acknowledgements

We thank professor Wenseleers and professor van Noort for their guidance during the research.

References

Breheny, P., & Huang, J. (2011). “Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection.” Annals of Applied Statistics, 5(1), 232–253.

Dunn, P. K., & Smyth, G. K. (2018). Generalized Linear Models With Examples in R. Springer Science+Business Media.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). “Regularization Paths for Generalized Linear Models via Coordinate Descent.” Journal of Statistical Software, 33(1), 1–22.

Frommlet, & Nuel, G. (2016). An adaptive ridge procedure for L0 regularization. PloS One, 11(2), e0148620–e0148620. https://doi.org/10.1371/journal.pone.0148620

Hardin, J. W., & Hilbe, J. M. (2018). Generalized Linear Models and Extensions, Fourth Edition. 598. Retrieved from https://www.stata-press.com/books/generalized-linear-models-and-extensions/

Hazimeh, H., & Mazumder, R. (2020). Fast best subset selection: Coordinate descent and local combinatorial optimization algorithms. Operations Research, 68(5), pp.1517–1537.

Jared, H. (2018). Ordinis:https://github.com/jaredhuling/ordinis

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). An Introduction to Statistical Learning (Springer texts in statistics). New York, NY: Springer.

Rubin, D. B. (2014). Iteratively Reweighted Least Squares. Wiley StatsRef: Statistics Reference Online, (3), 1–14. https://doi.org/10.1002/9781118445112.stat03199

Stellato, B., Banjac, G., Goulart, P., Bemporad, A., & Boyd, S. (2020). OSQP: an operator splitting solver for quadratic programs. Mathematical Programming Computation, 12(4), 637–672. https://doi.org/10.1007/s12532-020-00179-2

Stuetzle, W. (2001). The Conjugate Gradient Method.