Estimation of Sparse Generalized Linear Models: the dglars package

Luigi Augugliaro and Angelo M. Mineo

Abstract `dglars` is a public available `R` package that implements the method proposed in Augugliaro, Mineo and Wit (2013) developed to study the sparse structure of a generalized linear model. This method, called dgLARS, is based on a differential geometrical extension of the least angle regression method (LARS). The core of the `dglars` package consists of two algorithms implemented in `Fortran 90` to efficiently compute the solution curve; specifically a predictor-corrector algorithm and a cyclic coordinate descent algorithm.

Key words: generalized linear models, dgLARS, predictor-corrector algorithm, cyclic coordinate descent algorithm, sparse models, variable selection

1 Introduction

Nowadays, high-dimensional data sets, namely data sets where the number of predictors, say p, is larger than the sample size N, are becoming more and more common. Modern statistical methods developed to study this kind of data sets are usually based on the idea to use a penalty function to estimate a solution curve embedded in the parameter space and then to find the point that represents the best compromise between sparsity and predictive behaviour of the model. Recent statistical literature has a great number of contributions devoted to this problem: some important examples are the L_1-penalty function [8], the SCAD method [5] and the MC+ penalty function [9], among others.

Luigi Augugliaro
University of Palermo, Viale delle Scienze Ed. 13 e-mail: luigi.augugliaro@unipa.it

Angelo M. Mineo
University of Palermo, Viale delle Scienze Ed. 13 e-mail: angelo.mineo@unipa.it

Differently from the methods cited above, in [3] the authors propose a new approach based on the differential geometrical representation of a GLM. The derived method, that does not require an explicit penalty function, has been called differential geometric LARS (dgLARS) method because it is defined generalizing the geometrical ideas on which the least angle regression (LARS), proposed in [4], is based. Using the differential geometric characterization of the classical signed Rao score test statistic, dgLARS gains important theoretical properties that are not shared by other methods. From a computational point of view, the dgLARS method consists essentially in the computation of the implicitly defined solution curve. In [3] this problem is satisfactorily solved by using a predictor-corrector (PC) algorithm. In [2] is proposed a much more efficient cyclic coordinate descend (ccd) algorithm to fit the dgLARS solution curve when we work with an high-dimensional data set.

In this paper we present the `dglars` package that implements both the algorithms to compute the solution curve implicitly defined by dgLARS. The object returned by these functions is a S3 class object, for which specific methods and functions have been implemented. The package `dglars` is available under general public licence (GPL-2) from the Comprehensive `R` Archive Network at http://CRAN.R-project.org/package=dglars.

2 The `dglars` package

The `dglars` package is an `R` [6] package containing a collection of tools related to the dgLARS method. In the following of this section we describe the main functions available with this package. For a complete description of the all functions implemented in the `dglars` package the reader is referred to the manual of the package.

The function `dglars()` is the main function of the proposed package. It can be called with the following arguments

```
dglars(X, y, family = c("binomial", "poisson"), control = list())
```

where `X` is the design matrix of dimension $n \times p$, `y` is the n-dimensional response vector and `family` is the error distribution used in the model. Finally `control` is a named list of control parameters. For a complete description of this list, the interested readed is refereed to the corresponding help page.

To gain more insight on how to use the `dglars()` function we consider a simulated data set. We simulate a data set from a logistic regression model with sample size equal to 100 and $p = 5$ predictors. We also assume that only the first 2 predictors influence the response variable. The used `R` code is

```
R> n <- 100; p <- 4; s <- 2; X <- matrix(rnorm(n * p), n, p)
R> bs <- rep(1, s); Xs <- X[ , 1 : s]
R> eta <- drop(1 + drop(Xs %*% bs))
```

```
R> mu <- binomial()$linkinv(eta); y <- rbinom(n, 1, mu)
R> out_dglasso_pc <- dglars(X = X, y = y, family = "binomial")
```

`dglars()` returns a S3 class object

```
R> class(out_dglasso_pc)
[1] "dglars"
```

As shown in the following R code, the method function `print.dglars()` can be used to print out the basic information contained in a `dglars` object.

```
R> out_dglasso_pc

Call:  dglars(X = X, y = y, family = "binomial")

  Sequence        g    Dev     %Dev  df
       +x1  3.67566  134.6  0.00000   1
            3.06853  130.5  0.03080   2
       +x2  3.04937  130.3  0.03171   2
            0.21800  109.0  0.19047   3
       +x4  0.20859  109.0  0.19055   3
            0.05396  108.8  0.19188   4
       +x3  0.03199  108.8  0.19194   4
            0.00010  108.8  0.19198   5

Algorithm pc ( method = dgLASSO ) with exit = 0
```

The column `Sequence` shows that the dgLARS method first finds the true predictors and then includes the other false predictors. The column `g` reports the value of the parameter γ used in the dgLARS method to select the trade-off between sparsity of the estimated model and prediction behaviuor [3]. To be more specific, at the starting step the predictor `x1` makes the smallest angle with the tangent residual vector and then is included in the active set. The predictor `x2` is included in the active set at $\gamma^{(2)} = 3.04937$, this means that $\hat{\beta}_2(\gamma^{(2)})$ is equal to zero and then the number of non-zero estimated coefficients is equal to 2. The predictor `x4` is included at $\gamma^{(3)} = 0.20859$ and so on.

More information about the estimated sequence of models can be obtained using the method function `symmary.dglars()`. The output printed out by `summary.dglars()` is divided in three different sections. The first section completes the basic information printed out by `print.default()` showing the sequence of the Akaike Information Criterion (AIC) [1] and the sequence of the Bayesian Information Criterion (BIC) [7]. The ranking of the estimated models obtained by these two criteria are also showed and the corresponding best model is identified by an arrow. The last two sections show the estimated coefficients of the two best models. For the sake of brevity, the following R code shows only the first section printed out by `summary.dglars()`:

```
R> summary(out_dglasso_pc)

Call:  dglars(X = X, y = y, family = "binomial")

  Sequence        g    Dev  Complexity    AIC  Rank.AIC  Rank.BIC    BIC
       +x1  3.67566  134.6           1  136.6      8            6  139.2
            3.06853  130.5           2  134.5      7            8  139.7
       +x2  3.04937  130.3           2  134.3      6            7  139.5
            0.21800  109.0           3  115.0      2            2  122.8
       +x4  0.20859  109.0           3  115.0      1 <-      -> 1  122.8
            0.05396  108.8           4  116.8      4            4  127.2
       +x3  0.03199  108.8           4  116.8      3            3  127.2
            0.00010  108.8           5  118.8      5            5  131.8
```

3 Conclusion

In this paper we have described the `R` package `dglars`. This package implements the differential geometric extension of the LARS method proposed in [3] and called dgLARS. The use of this package is shown by means of a simulated data set. The output of the functions are presented in a way that is easy to interpret for people familiar with standard lm, glm or gam output.

References

1. Akaike H.: Information Theory as an Extension of the Maximum Likelihood Principle. In BN Petrov, F Czaki (eds.), Second International Symposium on Information Theory, pp. 267–281 (1973). Akademiai Kiado, Budapest.
2. Augugliaro L, Mineo AM, Wit E.C.: Differential Geometric LARS Via Cyclic Coordinate Descent Method. In *Proceedings of COMPSTAT 2012*, pp. 67–79 (2012). Limassol, Cyprus.
3. Augugliaro L, Mineo AM, Wit E.C.: Differential Geometric Least Angle Regression: A Differential Geometric Approach to Sparse Generalized Linear Models. J. R. Statist. Soc. B (2013).
4. Efron B, Hastie T, Johnstone I, Tibshirani R.: Least Angle Regression. Ann. Statist. **32**(2), 407–499 (2004).
5. Fan J, Li R.: Variable Selection Via Nonconcave Penalized Likelihood and Its Oracle Properties. J. Am. Statist. Ass. **96**(456), 1348–1360 (2001).
6. `R` Development Core Team (2012). `R`: A Language and Environment for Statistical Computing. `R` Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, http://www.R-project.org/.
7. Schwarz G.: Estimating the Dimension of a Model. Ann. Statist. **6**(2), 461–464 (1978).
8. Tibshirani R.: Regression Shrinkage and Selection Via the Lasso. J. R. Statist. Soc. B **58**(1), 267–288 (1996).
9. Zhang C.H.: Nearly Unbiased Variable Selection Under Minimax Concave Penalty. Ann. Statist. **38**(2), 894–942 (2010).