LASSO regression via smooth L_1-norm approximation

Vito M.R. Muggeo[1]

[1] Dip. Scienze Statistiche e Matematiche 'Vianelli', Università di Palermo

Abstract: This paper discusses estimation of regression model with LASSO penalty when the L_1-norm is replaced with its parametric smooth approximation. The resulting parameter estimators are more manageable than those from standard LASSO, standard errors are easy computed via a sandwich formula, and the model degrees of freedom may be computed straightforwardly. Moreover the resulting objective function may be minimized using usual optimization algorithms for regular models, for instance Newton-Raphson or iterative least squares.

Keywords: LASSO; L_1-norm; smooth models; least squares.

1 Introduction

The *Least Absolute Shrinkage and Selection Operator*, LASSO, was introduced by Tibshirani (1996) as a device to obtain sparse solution in linear regression models with a large number of covariates. 'Sparse solution' means that in the final solution some of the estimated β_js regression coefficients are automatically set to zero by the procedure, so allowing (continuous) variable selection and parameter estimation *simultaneously*. The key for this crucial and important feature is the L_1-norm penalty $\lambda \sum_j |\beta_j|$ which controls the sparseness of the solution via the tuning parameter $\lambda \geq 0$: at $\lambda = 0$ the solution corresponds to the OLS estimates (if these exists), and as λ increases more estimates are set to zero. For observed responses y_i and covariate vector $x_i \in R^p$, the L_1 penalized loss function for the regression model $\mu_i = x_i^T \beta$ may be written $\sum_i (y_i - x_i^T \beta)^2 + \lambda \sum_j |\beta_j|$.
Unfortunately LASSO does not come without concerns. Standard errors for the parameter estimates are not easily obtained in the L_1-penalized framework and some approximations have been discussed. Tibshirani (1996) proposed to use the ridge-type approximation $\sum_j |\beta_j| \approx \sum_j \beta_j^2 / |\tilde{\beta}_j|$ to justify a sandwich-type formula, where the tilde means an approximate known value. Osborne *et al.* (2000) showed that this approach breaks down as it provides $\mathrm{SE}(\hat{\beta}_j) = 0$ when $\hat{\beta}_j = 0$, which is inappropriate as also zero estimates would be associated with some degree of uncertainty; Osborne and co-workers derived a formula for covariance matrix which ensures positive standard errors for all coefficient estimates. A bootstrap approach was also

proposed (Tibshirani, 1996), but it fails to be consistent (Kyung et al., 2010) and moreover it may be troublesome in some contexts, especially in large dataset and/or complex models. Regardless of the approach used to compute standard errors, the sampling distribution of the regression parameter estimator with L_1 penalty is not multivariate Normal when at least one true coefficient is zero. Typically, the density of the sampling distribution with null parameter has positive probability mass at zero (Knight and Fu, 2000: Pötscher and Leeb, 2009; Kyung *et al.*, 2010). Non-normality of sampling distribution causes the standard errors to be somewhat useless for inference purpose, namely to build confidence intervals and to carry out hypothesis testing. Finally, from a practical viewpoint, although LASSO algorithms are well established in linear or even generalized linear models (Efron et al., 2004), their use may result not straightforward for more complex contexts, for instance models with heteroscedastic/autocorrelated errors and/or random effects.
In this paper we describe a smooth parametric approximation of the L_1-norm which allows to perform LASSO regression via iterative least squares and to get valid standard errors for all the parameter estimates.

2 Methods

We replace the usual L_1 penalty by its smooth and parametric approximation, which we call 'quasi-L_1', briefly qL_1; we refer the resulting qL_1 penalized regression as 'quasi-LASSO'. To begin with, we approximate the absolute value function,

$$|\beta| \approx Q(\beta) = -\beta + \frac{(\beta + c)^2}{2c} I(|\beta| \leq c) + 2\beta I(\beta > c), \tag{1}$$

where $I(\cdot)$ is the usual indicator function and c is a small known constant regulating the width of the bend around zero. Unlike $|\beta|$, $Q(\beta)$ is two-times differentiable at the origin for $c > 0$, and as $c \to 0$, $Q(\beta) \to |\beta|$. Using (1) the qL_1 penalty for a parameter vector is naturally given by $\sum_j |\beta_j| \approx \sum_j Q(\beta_j)$.
Figure 1 illustrates the LASSO and the quasi-LASSO penalties in the plane (β_1, β_2) for three different values of c; the smooth arc (for $c > 0$) replacing the kink in the neighbour of $\beta_j = 0$ guarantees differentiability of the objective function. The minimand loss function with the quasi-LASSO may be written as

$$\mathcal{L}_c = \sum_i (y_i - x_i^T \beta)^2 + \lambda \sum_j Q(\beta_j). \tag{2}$$

The main appealing of the function $Q(\cdot)$, and thus of qL_1, is that it admits first and second derivatives, and as a consequence, there exist the score and the hessian of the objective function (2). It can be shown that the

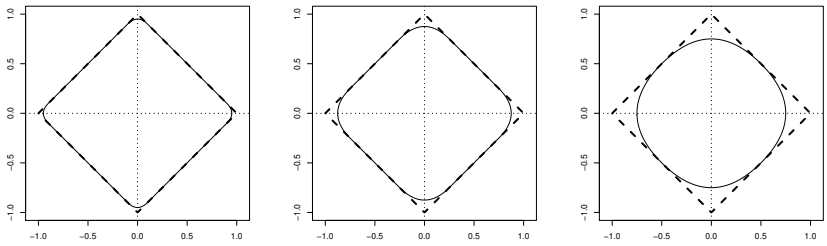

FIGURE 1. Contour plots of the LASSO (dashed lines) and the approximate LASSO (continuous lines) penalties for two coefficients and three different values of c.

Newton-Raphson step is equivalent to the least squares computation

$$\hat{\beta} = (X^T X + \lambda \widetilde{V})^{-1} X^T \tilde{z}^*, \tag{3}$$

with

$$\begin{aligned} z^* &= X^- X^T y + \lambda X^- (I_p 1_p - c\widetilde{V} 1_p - \widetilde{W} 1_p) \\ &= y + \lambda X^- (I_p 1_p - c\widetilde{V} 1_p - \widetilde{W} 1_p), \end{aligned} \tag{4}$$

where X is the usual $n \times p$ design matrix, X^- is the $n \times p$ Moore-Penrose generalized inverse of X^T, and y is response vector; moreover I_p is the $p \times p$ identity matrix and 1_p is a p-vector of ones, while $\widetilde{V} = c^{-1}\text{diag}(I(|\tilde{\beta}_j| \leq c))$ and $\widetilde{W} = \text{diag}(2I(\tilde{\beta}_j > c))$ are p-dimensional square matrices. Tilde values mean approximate values: at the first step, the initial guesses for the beta parameters may be obtained by standard OLS or even trivial starting values $(0, 0, \ldots, 0)^T$ may be employed; some empirical experience suggest that starting values are a minor issue and convergence is achieved in a few iterations. However whether several models have to be fitted (for instance to search for the optimal value of the tuning parameter) appropriate starting values may speed convergence overall. Note when additional covariates have to be included without penalizing corresponding coefficients (for instance the intercept), it suffices to include them in the design matrix X and to add zeroes to the main diagonals of the matrices V and W, such that these become blockdiag$(V, 0, \ldots, 0)$ and blockdiag$(W, 0, \ldots, 0)$.

Unlike the pure LASSO, owing to the parameterization of $Q(\beta)$ some estimates will never be exactly zero. However as the parameter c measures the amount of approximation at the kinks $\beta_j = 0$, in a 'model selection' context it is reasonable to consider zero the estimates fulfilling $|\hat{\beta}_j| \leq c$.

Parameter estimates depend on $c > 0$ and $\lambda \geq 0$ jointly. At $\lambda = 0$ the algorithm yields the OLS estimates regardless the value of c, while when $\lambda \neq 0$ the value of c is not trivial: if each $|\hat{\beta}_j| < c$, it is easy to see that $V = c^{-1} I_p$

and $W = 0$, hence $\tilde{z}^* = y$ and solution (3) reduces to the ridge penalty solution with tuning parameter λ/c. As sketched in the Introduction, it is possible to get asymptotic standard errors for the overall parameter vector estimate within the quasi-LASSO framework. The resulting sandwich formula is

$$\text{cov}(\hat{\beta}) = E[H]^{-1}\text{cov}(U)E[H]^{-T} \tag{5}$$

where H is the hessian matrix, $\text{cov}(U)$ is the covariance matrix of the gradient U, and '$-T$' means the transpose of the inverse. To apply the sandwich formula in practice, $E[H]$ is estimated by the hessian H evaluated at $\hat{\beta}$, and $\text{cov}(U) = \hat{\sigma}^2 X^T X$. The sandwich formula does not constrain to zero the standard errors for the (quasi) zero estimates and it works for any λ and c; moreover also note that the sampling distribution of the qL_1 estimates is Normal, so formula (5) may be applied to build valid confidence intervals even when the true parameter is zero.

3 Application and Simulation

We use the quasi-LASSO penalty to the well-known Prostate Cancer dataset analyzed by Tibshirani (1996). There are $n = 97$ subjects, $p = 8$ covariates (see Table 1) and the response variable is the log of prostrate specific antigen. Figure 2 illustrates the GCV and the BIC scores for the selection of the tuning parameter λ and the threshold c; to compute GCV and BIC we use the trace of the hat matrix XHX^T which is defined for the qL_1 penalty as the hessian H exists. Both GCV and BIC favour the smallest value of c which suggests sparsity in the solution; given $c = 0.1$ the BIC selects a somewhat more parsimonious model (i.e. larger λ). This is not surprising as it is well-known the BIC tends to select simple models (Zou et al., 2007).

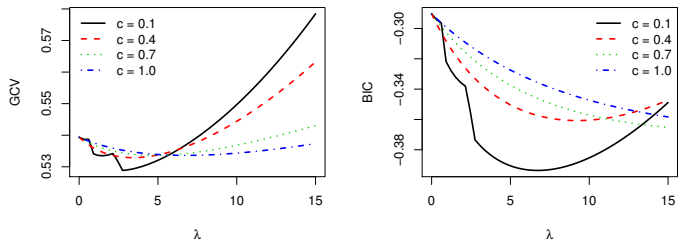

FIGURE 2. GCV and BIC scores with respect to λ for 4 values of c for the qL_1 regression for the Prostate Cancer dataset.

Table 1 shows the parameter estimates for the qL_1 regression. As reported in literature (e.g., Tibshirani, 1996), five out eight regressors appear to be not related to the response; however, unlike the L_1 penalty, the qL_1 framework provides confidence intervals which typically are quite informative.

TABLE 1. Parameter estimates, standard errors and Normal-based 95% confidence intervals via the qL_1 penalized regression ($c = 0.1$).

			95% CI	
covariate	Est	SE	inf	sup
lcavol	0.577	0.0927	0.395	0.759
lweight	0.227	0.0797	0.071	0.383
age	−0.063	0.0480	−0.157	0.031
lbph	0.083	0.0486	−0.013	0.178
svi	0.229	0.0894	0.054	0.404
lcp	0.003	0.0476	−0.090	0.097
gleason	0.036	0.0456	−0.054	0.125
pgg45	0.059	0.0448	−0.029	0.147

To illustrate the performance of the qL_1 we present results from a small simulation experiment ($n = 100$ with 2 predictors out $p = 20$ covariates). We consider the standard LASSO with tuning parameter selected by 10-fold CV and the qL_1 with λ selected by GCV using the trace of the hat matrix to quantify the model dimension. Figure 3 shows two sampling distributions for the estimator of a null parameter; unlike L_1, qL_1 appears to ensure Normal distribution for $\hat{\beta}$.

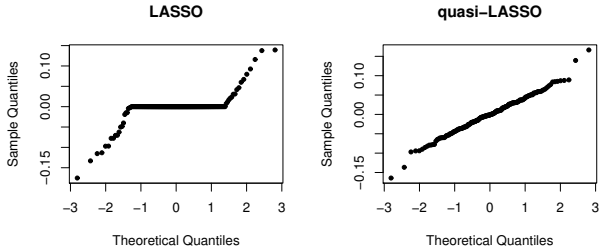

FIGURE 3. Normal Q-Q plots of sampling distributions of LASSO and quasi-LASSO estimator $\hat{\beta}$ when $\beta = 0$.

Table 2 shows the mean, standard deviation and average of the estimated SE for 4 parameter estimators; sandwich formula in the qL_1 framework appears to work reasonably well.

TABLE 2. Mean and standard deviation of the sampling distributions for 4 parameter estimators in the simulation study; $\bar{\text{SE}}$ indicates the averages of the standard errors obtained in each replicate (in each block the 4 columns refer respectively $\beta = 0.5, -0.5, 0, 0$).

	LASSO				quasi-LASSO			
m($\hat{\beta}$)	0.374	−0.358	0.000	0.001	0.390	-0.374	-0.001	0.003
sd($\hat{\beta}$)	0.116	0.100	0.036	0.036	0.109	0.096	0.051	0.050
$\bar{\text{SE}}$	0.103	0.103	0.107	0.108	0.099	0.098	0.048	0.047

4 Discussion

We have presented a smooth parametric approximation for the LASSO penalty. The resulting quasi-LASSO regression may be implemented via iterative least squares and guarantees normality of the parameter estimator, even in presence of zero coefficients. A sandwich formula is available and allows to build reliable normal-based confidence intervals. Smooth approximations of the L_1-norm have been discussed by Osborne *et al.* (2000); however their approximations, different from the (1), are employed to motivate some formulas for the covariance matrix and not for estimation. The proposed qL_1 approximation depends upon a threshold value $c > 0$. As c increases qL_1 reduces to L_2 (i.e. ridge regression), and 'sparsity' is lost. Therefore selection of c is an important issue of the proposed approach which deserves major investigation. Another important issue refers to the computation of the model *df* which may be computed by the number of 'non-zero' estimates, namely $\#\{|\hat{\beta}_j| > c\}$, or alternatively it is possible use the trace of the 'hat matrix', $df = \mathrm{tr}(X^T X H^{-1})$. In conclusion, we believe the proposed approach represents a possible alternative to get reliable inference (confidence intervals and p-value) within the frequentist framework as an alternative to the Bayesian one (Kyung et al., 2010).

References

Knight K. and Fu W. (2000) Asymptotics for lasso-type estimators. *Annals of Statistics*, **28**, 1356-1378.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004) Least angle regression. *Annals of Statistics*, **32**, 407–489.

Kyung, M., Gilly, J., Ghoshz, M.,and Casella, G. (2010) Penalized Regression, Standard Errors, and Bayesian Lassos. *Bayesian Analysis*, **5**, 1–44.

Osborne, M. R., Presnell, B., and Turlach, B. A. (2000). On the LASSO and its dual. *Journal of Computational Graphical Statistics*, **9**, 319–337.

Pötscher, B. M., and H. Leeb (2009) On the distribution of penalized maximum likelihood estimators: The LASSO, SCAD, and thresholding. *Journal of Multivariate Analysis*, **10**, 2065–2082.

Tibshirani (1996) Regression shrinkage and selection via the lasso. *Journal of Royal Statistical Society B*, **58**, 267–288.

Zou, H., Hastie, T., and Tibshirani, R. (2007) On the 'degrees of freedom' of the lasso. *Annals of Statistics*, **35**, 2173–2192.