A Sensitivity Analysis and Error Bounds for the Adaptive Lasso

Tathagata Basu[1], Jochen Einbeck[1], Matthias C. M. Troffaes[1]

[1] Durham University, United Kingdom

E-mail for correspondence: tathagata.basu@durham.ac.uk

Abstract: Sparse regression is an efficient statistical modelling technique which is of major relevance for high dimensional problems. There are several ways of achieving sparse regression, the well-known lasso being one of them. However, lasso variable selection may not be consistent in selecting the true sparse model. Zou (2006) proposed an adaptive form of the lasso which overcomes this issue, and showed that data driven weights on the penalty term will result in a consistent variable selection procedure. Weights can be informed by a prior execution of least squares or ridge regression. Using a power parameter on the weights, we carry out a sensitivity analysis for this parameter, and derive novel error bounds for the Adaptive lasso.

Keywords: Adaptive lasso; Sensitivity analysis; Variable selection.

1 Introduction

Let $\mathbf{X} = (X_1, \cdots, X_p)$ with $X_j = (X_{1j}, \dots X_{nj})^{\mathrm{T}}$ for $1 \leq j \leq p$, and $Y = (Y_1, \dots, Y_n)^{\mathrm{T}}$. We can characterise their relation in the linear regression setting

$$Y = \mathbf{X}\boldsymbol{\beta} + \epsilon, \tag{1}$$

where $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^{\mathrm{T}}$ is a vector of regression coefficients and $\epsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I}_n)$, with $\mathbf{I}_n$ denoting the n-dimensional identity matrix. We assume $\mathbf{X}$ and Y to be scaled to mean 0.

The least squares method is the conventional way to estimate these regression coefficients. However, in high dimension (i.e $p > n$), the least squares method, which involves inversion of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$, cannot be used. Several estimators have been proposed which solve the issue by introducing bias in the estimation process. Tikhonov (1963) introduced ℓ_2 penalised regression or Ridge regression. The ℓ_2 penalty achieves a stable solution through the

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

eigen value decay method, which, however, fails to be sparse which is a desirable property in high dimensional statistics. Tibshirani (1996) introduced the lasso or least absolute shrinkage and selection operator, which attains sparsity through a ℓ_1 penalty. Zou (2006) proposed an adaptive form of lasso based on data-driven weights in the penalty term that satisfies desired asymptotic properties for high-dimensional problems as suggested by Fan and Li (2001). We exploit the framework given by Zou (2006) to investigate and understand the sensitivity of the adaptive lasso. For this we apply a two-step approach. We employ least squares or ridge estimates, say $\hat{\beta}_j$, and a parameter γ to initialise the weights of type $1/|\hat{\beta}_j|^\gamma$ which are then embedded in the penalty term. The effect of the parameter γ is then investigated, theoretically, through error bounds, and experimentally, through a sensitivity analysis.

2 Adaptive Lasso

Let us consider the linear model (1) which can be written in alternative form as

$$\mathbb{E}[Y \mid \mathbf{X}] = \mathbf{X}\boldsymbol{\beta} = \beta_1 X_1 + \cdots + \beta_p X_p. \tag{2}$$

Note that $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is guaranteed to be positive semi-definite but not necessarily positive definite, even for $p < n$. We make the following two assumptions on the design $\mathbf{X}$:

(A1) $\mathbb{E}[\mathbf{X}^{\mathrm{T}}\epsilon \mid \mathbf{X}] = 0$

(A2) $\lim_{n\to\infty} \frac{1}{n}\mathbf{X}^{\mathrm{T}}\mathbf{X} = \Sigma$ exists, where Σ is positive definite.

Let $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1, \cdots, \hat{\boldsymbol{\beta}}_p)^{\mathrm{T}}$ be any root-n consistent estimator of $\boldsymbol{\beta}$. Then the adaptive lasso estimates are given by

$$\hat{\boldsymbol{\beta}}_{\mathrm{alasso}}(\lambda, \gamma) = \arg\min_{\boldsymbol{\beta}} \left(\frac{1}{2}\|Y - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{j=1}^{p} w_j(\gamma)|\beta_j| \right) \tag{3}$$

where

$$w_j(\gamma) = |\hat{\beta}_j|^{-\gamma}, \quad \text{for} \quad \gamma > 0. \tag{4}$$

We generally use least squares estimates or ridge estimates as weights since these are root-n consistent.

3 Main Result

Let $\hat{\boldsymbol{\beta}}_{\mathrm{alasso}}(\lambda, \gamma)$ be the adaptive lasso estimates with respect to the parameters λ and γ and $\Sigma_n = \frac{1}{n}\mathbf{X}^{\mathrm{T}}\mathbf{X}$. Let $\boldsymbol{\beta}^*$ be the true regression coefficients.

Theorem: For any root-n consistent estimate $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \cdots, \hat{\beta}_p)$, we have the following error bounds:

$$\left\| \hat{\boldsymbol{\beta}}_{\text{alasso}}(\lambda, \gamma) - \boldsymbol{\beta}^* \right\|_2^2 \leq \frac{\sigma^2}{n} \left\| \Sigma_n^{-1} \right\| + \frac{\lambda^2 p}{n^2} \left\| \Sigma_n^{-1} \right\|^2 \min_{1 \leq j \leq p} |\hat{\beta}_j|^{-2\gamma} \quad (5)$$

$$\left\| Y - \mathbf{X}\hat{\boldsymbol{\beta}}_{\text{alasso}}(\lambda, \gamma) \right\|_2^2 \leq \frac{\lambda^2 p}{n} \left\| \Sigma_n^{-1} \right\| \min_{1 \leq j \leq p} |\hat{\beta}_j|^{-2\gamma} \quad (6)$$

We see that the error bounds increase with increasing λ (increased bias from regularisation) but tend to decrease with increasing γ.

4 Simulation Study

We simulate the predictors from a standard normal distribution such that, $X_{ij} \sim N(0, 1)$ for $j = 1, \cdots, 20$ and $i = 1, \cdots, n$. We assign the regression coefficients to be $(\beta_1, \cdots, \beta_6) = (5, 3, 1, -1, -3, -5)$ and $\beta_j = 0$ for $j > 6$. We consider standard normal noise to construct the response vector $y_i = \sum_{j=1}^{6} X_{ij}\beta_j + \epsilon_i$ where, $\epsilon_i \sim N(0, 1)$ for $i = 1, \cdots, n$. The experiment is repeated for $n = 100, 500, 1000$.
We analyse the sensitivity of the model for $0 \leq \gamma \leq 1$ ($\gamma = 0$ yields regular lasso estimates). We use least squares estimates for the choice of weights. In Table 1, we compare prediction accuracy of different lasso variants, and also display the number of active co-variates, p^*. In the first row we give the results of the adaptive lasso for $\gamma = 1$. We specify λ through cross-validation. In the next three rows we show results for varying γ and fixed λ. In Figure 1, we show the coefficient path and RMSE curve evaluated over γ for 100 observations. From Figure 1 we see that as the value of γ increases, the bias and RMSE decrease which is plausible in the light of Theorem 1. However, we also notice that it overfits and selects six extra variables as important. In the last row we show results from the lasso.

5 Conclusion

We have presented a sensitivity analysis for the adaptive lasso with respect to γ, and obtained novel bounds for the lasso estimates. We have shown through simulation that the bias due to regularisation with λ can be reduced for larger values of γ, however, especially for small sample sizes, at the potential expense of overfitting and selection of some non-important variables in the model.

Acknowledgments: This work is funded by the European Commission's H2020 programme, through the UTOPIAE Marie Curie Innovative Training Network, H2020-MSCA-ITN-2016, Grant Agreement number 722734.

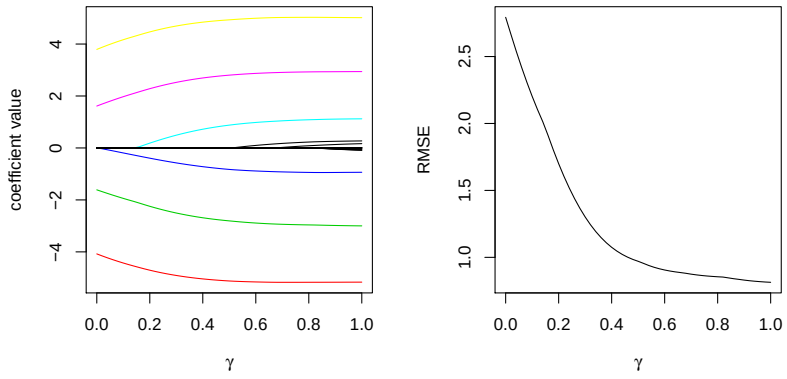

FIGURE 1. Coefficient path and fitting accuracy w.r.t. γ ($\lambda = 1$) for $n = 100$.

TABLE 1. Comparison of prediction accuracy (RMSE) between different methods.

	$n = 100$		$n = 500$		$n = 1000$	
	RMSE	p^*	RMSE	p^*	RMSE	p^*
Adaptive Lasso						
$\gamma = 1$, λ by CV	0.94	6	1.02	6	0.99	6
$\gamma = 0.1$, $\lambda = 1$	2.20	5	2.02	6	1.94	6
$\gamma = 0.5$, $\lambda = 1$	0.97	6	1.05	6	1.02	6
$\gamma = 1$, $\lambda = 1$	0.81	12	0.99	6	0.97	6
Lasso, λ by CV	0.93	10	1.03	6	1.00	6

References

Fan, J., and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96**(456), 1348 – 1360.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **58**(1), 267 – 288.

Tikhonov, A. N. (1963). On the solution of ill-posed problems and the method of regularization. *Dokl. Akad. Nauk SSSR*, **151**(3), 501 – 504.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, **101**(476), 1418 – 1429.