MATH 680: Computation Intensive Statistics **Winter 2018**

Lecture 19: March 21

Lecturer: Lecturer: Yi Yang *Scribes: Zafarali Ahmed, Ismaila Diedhiou Balde, David Fleischer*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

19.1 Lasso

19.1.1 Two Source of Inspiration for the Lasso

1. Non-negative Garrote [B95]: The non-negative Garrote is defined by a multiple-step process. First, solve the ordinary least-squares regression problem

$$\widehat{\beta}^{\text{ols}} = \arg\min_{\beta \in \mathbb{R}^p} \|\mathbf{y} - \mathbf{X}\beta\|_2^2,$$

for $\widehat{\beta}^{\text{ols}} = \left(\widehat{\beta}_1^{\text{ols}}, ..., \widehat{\beta}_p^{\text{ols}}\right)^T$. Next, use the OLS estimate to define the (non-negative Garrote) constrained minimization problem (with respect to c_j)

$$\text{Minimize} \quad \|\mathbf{y} - \sum_{j=1}^p \mathbf{X}_j \widehat{\beta}_j^{\text{ols}} c_j\|_2^2 \quad \text{subject to}$$

$$\sum_{j=1}^p c_j \leq B \quad \text{and} \quad c_j \geq 0 \;\forall\, j.$$

where $B \in \mathbb{R}^+$ is a tuning parameter similar to λ in the unconstrained Lasso problem, or t in the constrained Lasso problem. Then, the non-negative Garrote solution is defined by scaling the OLS solution by the solutions $\widehat{c}_j$ to the above minimization problem. In particular,

$$\widehat{\beta}_j^{\text{garrote}} = \widehat{c}_j \widehat{\beta}_j^{\text{ols}}.$$

2. Wavelet shrinkage [DJ95]: Set $y_i = \mu_i + \epsilon_i$, for $\epsilon_i \sim \mathcal{N}(0,1)$, $i = 1, ..., n$. Then, define the estimate $\widehat{\mu}_i$ of μ_i as the solution to the unconstrained minimization problem

$$\widehat{\mu}_i = \arg\min_{\mu} \left\{ \frac{1}{2} \left(y_i - \mu\right)^2 + \sqrt{2\log n}|\mu| \right\}$$

19.1.1.1 Non-Negative Garrote (Breiman, 1995, Technometrics)

"Much work and research have gone into subset selection regression, but the basic method remains flawed by its relative lack of accuracy and instability. Subset regression either zeros a coefficient, it is not in the selected subsets, or inflates it. Ridge regression gains its accuracy by selective shrinking."

"Methods that select subsets, are stable, and shrink are needed."

The garrote eliminates some variables, shrinks others, and is relatively stable (compared with the subset selection algorithm).

The non-negative garrote depends on both the sign and the magnitude of the OLS estimates. OLS estimates may behave poorly in some settings, such as overfit or highly correlated covariates. The non-negative garrote may suffer as a result.

Tibshirani proposed the LASSO with the goal to avoid the explicit use of the OLS estimates, as used in the non-negative garrote algorithm.

19.1.2 Computing the Lasso Estimator

- Use a standard quadratic program solver [T96].
- Shooting algorithm [F98].
- Homotopy method [OPT00].
- Least Angle Regression and the LARS algorithm [EHJT04], `R` package: `lars`.

  > Tibshirani's Lasso algorithm has had little impact on statistical practice. Two particular reasosn for this may be the relative inefficiency of the original Lasso algorithm, and the relative complexity of more recent Lasso algorithms (e.g., Osborn et al., 2000). [MR04].

- Coordinate descent [FHT10], `R` package: `glmnet`.
- Generalized coordinate descent [YZ15] `R` package: `gcdnet`.

19.1.2.1 Comments on the Lasso Estimator

The Lasso estimator $\widehat{\boldsymbol{\beta}}$ is given by the solution to the unconstrained minimization problem

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

Note that

- The solution $\widehat{\boldsymbol{\beta}}$ varies continuously with the λ-convexity of the objective function.
- There are a sequence of λ's which we call *transition points*, and the support of $\widehat{\boldsymbol{\beta}}$ stays locally constant between two adjacent transition points.
- The signs of the non-zero elements of $\widehat{\boldsymbol{\beta}}$ stays locally constant when λ is between two transition points.
- For any given λ, the probability that λ is a transition point is zero. That is, the set of transition points along $\mathbb{R}^+$ is of measure zero.

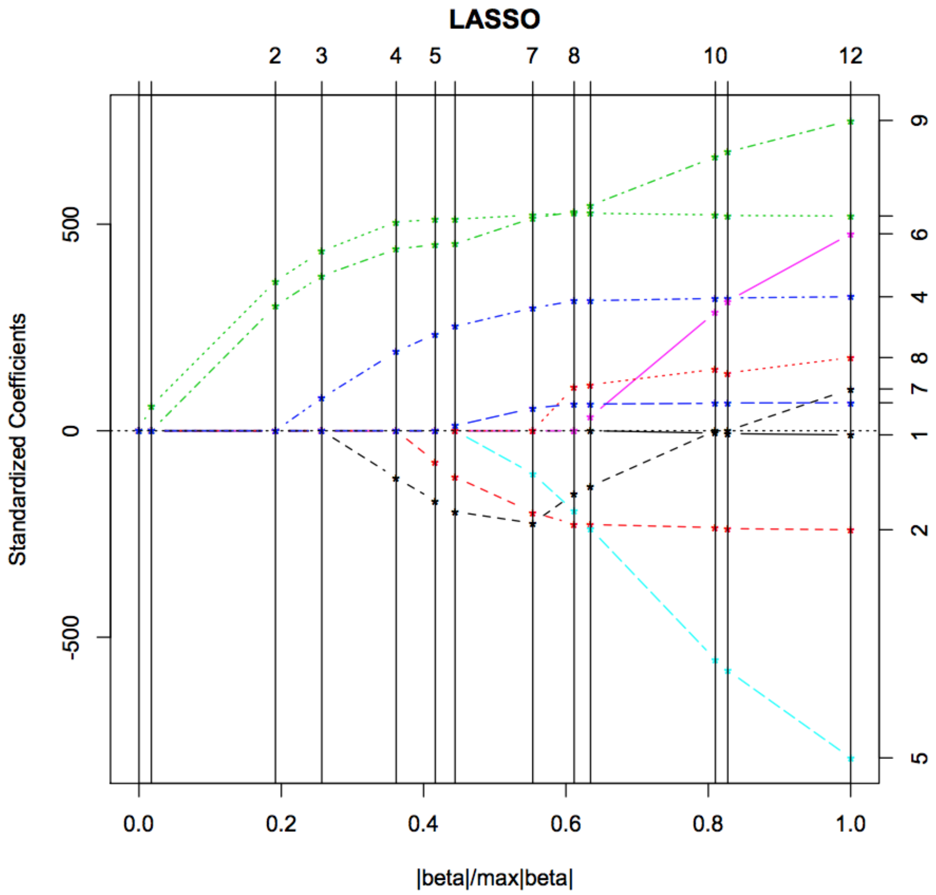

Figure 19.1: Lasso coefficient solution path.

19.1.3 Theoretical Considerations for the Lasso Estimator

Define A to be the *active set* of coefficients

$$A = \left\{ j \,:\, \widehat{\boldsymbol{\beta}} \neq 0 \right\},$$

and S_A to be the *signs* of the elements of A

$$\begin{aligned} S_A &= (s_j)\,,\ j \in A, \quad \text{for} \\ s_j &= \text{sign}\left(\widehat{\boldsymbol{\beta}}_j\right). \end{aligned}$$

Then, write $\widehat{\boldsymbol{\beta}} = \left(\widehat{\boldsymbol{\beta}}_A.,0\right)$. The solution $\widehat{\boldsymbol{\beta}}_A$ satisfies

$$-\mathbf{X}_A^T\left(\mathbf{y} - \mathbf{X}_A\widehat{\boldsymbol{\beta}}_A\right) + \lambda S_A = 0$$

which implies

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_A &= \left(\mathbf{X}_A^T\mathbf{X}_A\right)^{-1}\left(\mathbf{X}_A\mathbf{y} - \lambda S_A\right)\\ &= \left(\mathbf{X}_A^T\mathbf{X}_A\right)^{-1}\mathbf{X}_A\mathbf{y} - \lambda\left(\mathbf{X}_A^T\mathbf{X}_A\right)^{-1}S_A\\ \Longrightarrow \frac{d\widehat{\boldsymbol{\beta}}_A}{d\lambda} &= -\left(\mathbf{X}_A^T\mathbf{X}_A\right)^{-1}S_A.\end{aligned}$$

The above relation holds over the interval between two transition points since the solution path of $\widehat{\boldsymbol{beta}}$ is continuous between transition points. This result implies that the Lasso solution paths are piecewise linear (piecewise between transition points).

Next, the residuals $\mathbf{y} - \widehat{\mathbf{y}}$ are given by

$$\mathbf{y} - \widehat{\mathbf{y}} = \mathbf{y} - \mathbf{X}_A\left(\mathbf{X}_A^T\mathbf{X}_A\right)^{-1}\mathbf{X}_A^T\mathbf{y} + \lambda\mathbf{X}_A\left(\mathbf{X}_A^T\mathbf{X}_A\right)^{-1}S_A.$$

Thus, taking the derivative of the residuals $\mathbf{y} - \widehat{\mathbf{y}}$ with respect to λ yields

$$\begin{aligned}\frac{d\left(\mathbf{y} - \widehat{\mathbf{y}}\right)}{d\lambda} &= \mathbf{X}_A\left(\mathbf{X}_A^T\mathbf{X}_A\right)^{-1}S_A\\ &= V_A.\end{aligned}$$

Note that V_A is a special vector because

$$\begin{aligned}\mathbf{X}_A^T V_A &= \mathbf{X}_A^T\mathbf{X}_A\left(\mathbf{X}_A^T\mathbf{X}_A\right)^{-1}S_A\\ &= S_A.\end{aligned}$$

That is, $\mathbf{X}_A^T$ projects V_A to the signs of the active set, S_A, i.e., for each $j \in A$, the inner product between X_j and V_A is either $+1$ or -1 (depending on the sign of the corresponding solution $\widehat{\beta}_j$). Note that V_A was the derivative of the residual vector with respect to λ, and so the residual vector moves along a direction with equal angle with respect to all active covariates.

References

[B95] L. Breiman,“Better subset regression using the nonnegative garrote,” *Technometrics*, 37.4, 1995, pp. 373-384.

[DJ95] D.L. Donoho and I.M. Johnstone,“Adaptive to unknown smoothness via wavelet shrinkage,” *Journal of the american statistical association*, 90.432, 1995, pp. 1200-1224.

[T96] R. Tibshirani,“Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society. Series B (Methodological)*, 1996, pp. 267-288.

[F98] W.J. Fu,“Penalized regressions: the bridge versus the lasso,” *Journal of the computational and graphical statistics*, 7.3, 1998, pp. 397-416.

[OPB00] M.R. Osborne and B. Presnell and B.A. Turlach,“A new approach to variable selection in least squares,” *IMA journal of numerical analysis*, 20.3, 2000, pp. 389-403.

[EHJT04] B. Efron and T. Hastie and I. Johnston and R. Tibshirani,“Least angle regression,” *The Annals of Statistics*, 32.2, pp. 407-499.

[MR04] D. Madigan and G.Ridgeway,"[Least Angle Regression]: Discussion," *The Annals of Statistics*, 32.2, 2004, pp. 465-469.

[FHT10] J. Friedman and T. Hastie and R. Tibshirani,"Regularization paths for generalized linear models via coordinate descent",*Journal of statistical software*, 33.1, 2010, pp. 1.

[YZ15] Y. Yang and H. Zou,"A Fast Unified Algorithm for Solving Group-Lasso Penalized Learning Problems",*Statistics and Computing*, 25.6, 2015, pp. 1129-141.