Shrinkage methods: Least angles regression and Lasso

Nagarajan Natarajan, vishvAs vAsuki

October 7, 2009

Outline

1. Outline
2. Least Angle regression
3. LAR and Lasso: the connection
   - Lasso
   - The connection
4. Conclusion

The problem and the notation

The problem and the notation

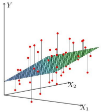


FIGURE 3.1. *Linear least squares fitting with $X \in \mathbb{R}^2$. We seek the linear function of X that minimizes the sum of squared residuals from Y.*

- Given N data points.

The problem and the notation

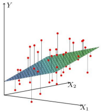


FIGURE 3.1. *Linear least squares fitting with $X \in \mathbb{R}^2$. We seek the linear function of X that minimizes the sum of squared residuals from Y.*

- Given N data points.
- Arrange them as rows in matrix X. $\begin{pmatrix} x_1 \\ x_2 \\ .. \end{pmatrix}$.

The problem and the notation

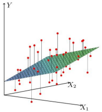


FIGURE 3.1. *Linear least squares fitting with $X \in \mathbb{R}^2$. We seek the linear function of X that minimizes the sum of squared residuals from Y.*

- Given N data points.
- Arrange them as rows in matrix X. $\begin{pmatrix} x_1 \\ x_2 \\ .. \end{pmatrix}$.
- Columns of X, y are centered.

The problem and the notation

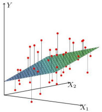


FIGURE 3.1. *Linear least squares fitting with $X \in \mathbb{R}^2$. We seek the linear function of X that minimizes the sum of squared residuals from Y.*

- Given N data points.
- Arrange them as rows in matrix X. $\begin{pmatrix} x_1 \\ x_2 \\ .. \end{pmatrix}$.
- Columns of X, y are centered.
- Solve: $X\beta \approx y$.

The problem and the notation

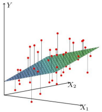


FIGURE 3.1. *Linear least squares fitting with $X \in \mathbb{R}^2$. We seek the linear function of X that minimizes the sum of squared residuals from Y.*

- Given N data points.
- Arrange them as rows in matrix X. $\begin{pmatrix} x_1 \\ x_2 \\ .. \end{pmatrix}$.
- Columns of X, y are centered.
- Solve: $X\beta \approx y$.
- Assume feature vectors have norm 1. $XDD^{-1}\beta \approx y$

Shrinkage methods

Shrinkage methods

- Ridge regression.

Shrinkage methods

- Ridge regression.
- Lasso.

Shrinkage methods

- Ridge regression.
- Lasso.
- Other penalties.

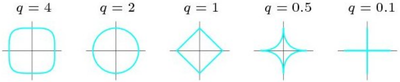


FIGURE 3.12. *Contours of constant value of* $\sum_j |\beta_j|^q$ *for given values of* q.

Shrinkage methods

- Ridge regression.
- Lasso.
- Other penalties.

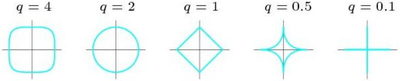


FIGURE 3.12. *Contours of constant value of $\sum_j |\beta_j|^q$ for given values of q.*

- Least angle regression.

From: [2], [1].

Outline

1 Outline

2 Least Angle regression

3 LAR and Lasso: the connection
- Lasso
- The connection

4 Conclusion

Forward selection

Forward selection

- Want sparse solution.

Forward selection

- Want sparse solution.
- How to balance love of sparsity with desire for a good fit?

Classical Forward Stepwise Selection

Classical Forward Stepwise Selection

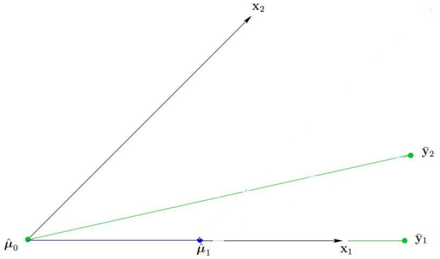
- Init: $\beta = 0, A = \phi$.

Classical Forward Stepwise Selection

- Init: $\beta = 0, A = \phi$.
- Current fit: $\hat{\mu} = X\beta$. Residue: $r = y - \hat{\mu}$. Current set of features: A.

Classical Forward Stepwise Selection

- Init: $\beta = 0, A = \phi$.
- Current fit: $\hat{\mu} = X\beta$. Residue: $r = y - \hat{\mu}$. Current set of features: A.
- Grow A one feature at a time: Select $x_j = argmax_i\{\langle x_i, r\rangle\}$.



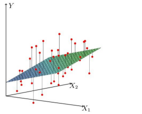
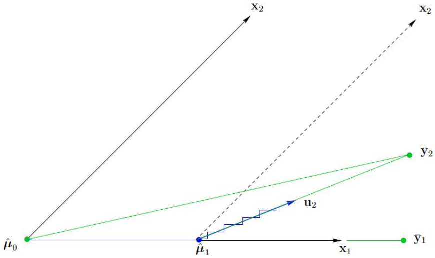


FIGURE 3.1. *Linear least squares fitting with $X \in \mathbb{R}^2$. We seek the linear function of X that minimizes the sum of squared residuals from Y.*

LAR Algorithm



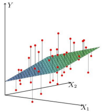
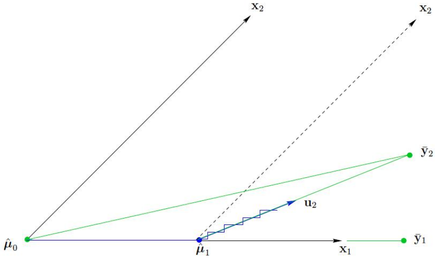


FIGURE 3.1. *Linear least squares fitting with $X \in \mathbb{R}^2$. We seek the linear function of X that minimizes the sum of squared residuals from Y.*

LAR Algorithm



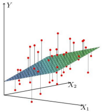
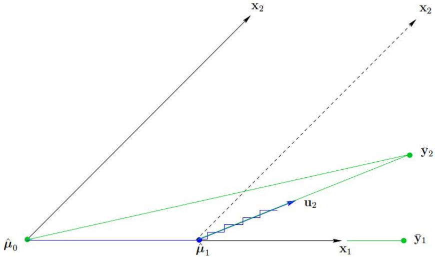


FIGURE 3.1. *Linear least squares fitting with $X \in \mathbb{R}^2$. We seek the linear function of X that minimizes the sum of squared residuals from Y.*

- Project residue r on subspace(A), get u_i.

LAR Algorithm



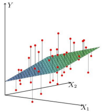
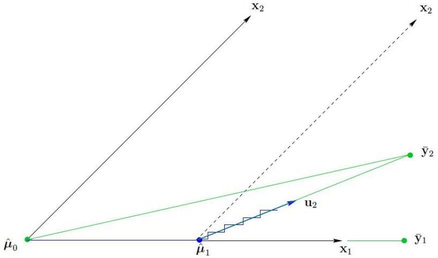


FIGURE 3.1. *Linear least squares fitting with $X \in \mathbb{R}^2$. We seek the linear function of X that minimizes the sum of squared residuals from Y.*

- Project residue r on subspace(A), get u_i.
- Set $\beta_A = \beta_A + \gamma_i$.
  So, $\mu = X\beta$ increases along u_i
  until you find $x_k : \langle x_k, r\rangle = \langle x_j, r\rangle$.

LAR Algorithm



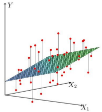
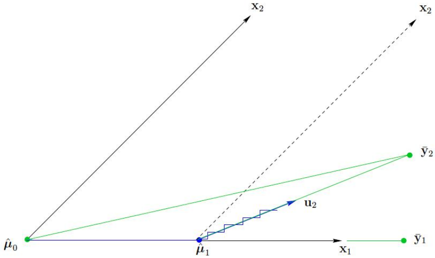


FIGURE 3.1. *Linear least squares fitting with $X \in \mathbb{R}^2$. We seek the linear function of X that minimizes the sum of squared residuals from Y.*

- Project residue r on subspace(A), get u_i.
- Set $\beta_A = \beta_A + \gamma_i$.
  So, $\mu = X\beta$ increases along u_i
  until you find $x_k : \langle x_k, r \rangle = \langle x_j, r \rangle$.
- $A = A \cup \{x_k\}$.

LAR Algorithm



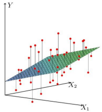


FIGURE 3.1. Linear least squares fitting with $X \in \mathbb{R}^2$. We seek the linear function of X that minimizes the sum of squared residuals from Y.

- Project residue r on subspace(A), get u_i.
- Set $\beta_A = \beta_A + \gamma_i$.
  So, $\mu = X\beta$ increases along u_i
  until you find $x_k : \langle x_k, r\rangle = \langle x_j, r\rangle$.
- $A = A \cup \{x_k\}$.
- Note! At any point, the residue makes same angle with all $x_i \in A$.

See how the correlation evolves

See how the correlation evolves

- $\hat{c}(\beta)$: the vector of correlations of the residue with $\{x_i\}$

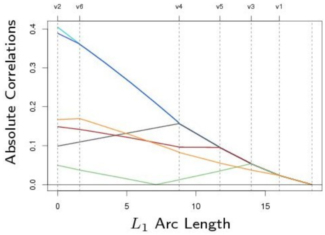


FIGURE 3.14. *Progression of the absolute correlations during each step of the LAR procedure, using a simulated data set with six predictors. The labels at the top of the plot indicate which variables enter the active set at each step. The step length are measured in units of L_1 arc length.*

Improvement over forward stepwise (greedy)

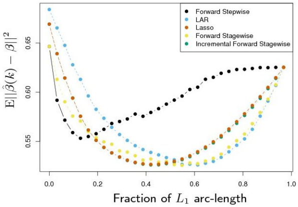


FIGURE 3.16. *Comparison of LAR and lasso with forward stepwise, forward stagewise (FS) and incremental forward stagewise (FS_0) regression. The setup*

How to balance love of sparsity with desire for a good fit?

Outline

1. Outline
2. Least Angle regression
3. LAR and Lasso: the connection
   - Lasso
   - The connection
4. Conclusion

Outline

1. Outline
2. Least Angle regression
3. LAR and Lasso: the connection
   - Lasso
   - The connection
4. Conclusion

The objective

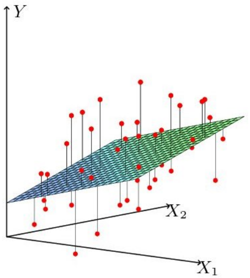


FIGURE 3.1. *Linear least squares fitting with $X \in \mathbb{R}^2$. We seek the linear function of X that minimizes the sum of squared residuals from Y.*

The objective

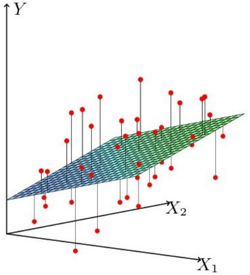


FIGURE 3.1. *Linear least squares fitting with $X \in \mathbb{R}^2$. We seek the linear function of X that minimizes the sum of squared residuals from Y.*

- $\hat{\beta} = argmin_{\beta} \|y - X\beta\|_2^2$ subject to $\sum |\beta_i| \leq t$.

The objective

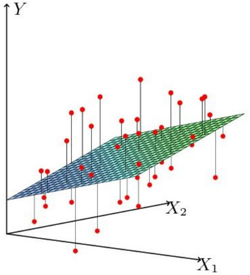


FIGURE 3.1. *Linear least squares fitting with $X \in \mathbb{R}^2$. We seek the linear function of X that minimizes the sum of squared residuals from Y.*

- $\hat{\beta} = argmin_{\beta} \|y - X\beta\|_2^2$ subject to $\sum |\beta_i| \leq t$.
- Same as $f(\hat{\beta}) = \min_{\beta} \|y - X\beta\|_2^2 + \lambda \sum_i |\beta_i|$.

Lasso for sparsity

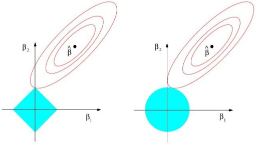


FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions* $|\beta_1| + |\beta_2| \le t$ *and* $\beta_1^2 + \beta_2^2 \le t^2$, *respectively, while the red ellipses are the contours of the least squares error function.*

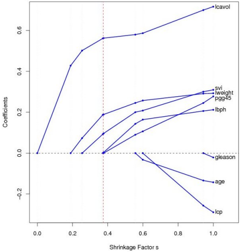


FIGURE 3.10. *Profiles of lasso coefficients, as the tuning parameter t is varied. Coefficients are plotted versus* $s = t/\sum_1^p |\hat{\beta}_j|$. *A vertical line is drawn at* $s = 0.36$, *the value chosen by cross-validation. Compare Figure 3.8 on page 9; the lasso profiles hit zero, while those for ridge do not. The profiles are piece-wise linear, and so are computed only at the points displayed;*

$$\hat{\beta} = argmin_{\beta} \|y - X\beta\|_2^2 \text{ subject to } \sum |\beta_i| \le t.$$

Lasso for sparsity

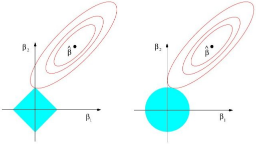


FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions* $|\beta_1| + |\beta_2| \le t$ *and* $\beta_1^2 + \beta_2^2 \le t^2$, *respectively, while the red ellipses are the contours of the least squares error function.*

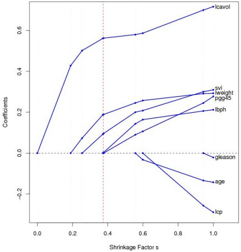


FIGURE 3.10. *Profiles of lasso coefficients, as the tuning parameter t is varied. Coefficients are plotted versus* $s = t/\sum_1^p |\hat{\beta}_j|$. *A vertical line is drawn at* $s = 0.36$, *the value chosen by cross-validation. Compare Figure 3.8 on page 9; the lasso profiles hit zero, while those for ridge do not. The profiles are piece-wise linear, and so are computed only at the points displayed;*

$\hat{\beta} = argmin_{\beta} \|y - X\beta\|_2^2$ subject to $\sum |\beta_i| \le t$.

- Quiz: What t will reduce the problem to least squares?

Lasso for sparsity

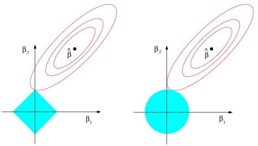


FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions* $|\beta_1| + |\beta_2| \leq t$ *and* $\beta_1^2 + \beta_2^2 \leq t^2$*, respectively, while the red ellipses are the contours of the least squares error function.*

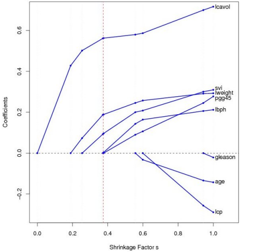


FIGURE 3.10. *Profiles of lasso coefficients, as the tuning parameter t is varied. Coefficients are plotted versus* $s = t/\sum_1^p |\hat{\beta}_j|$*. A vertical line is drawn at* $s = 0.36$*, the value chosen by cross-validation. Compare Figure 3.8 on page 9; the lasso profiles hit zero, while those for ridge do not. The profiles are piece-wise linear, and so are computed only at the points displayed;*

$\hat{\beta} = argmin_{\beta} \|y - X\beta\|_2^2$ subject to $\sum |\beta_i| \leq t$.

- Quiz: What t will reduce the problem to least squares?
- $t = \sum |\beta_i^*|$.

Outline

1. Outline
2. Least Angle regression
3. LAR and Lasso: the connection
   - Lasso
   - The connection
4. Conclusion

LARS: Solving LAR with Lasso

LARS: Solving LAR with Lasso

- Experimental observation

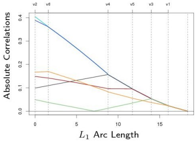


FIGURE 3.14. *Progression of the absolute correlations during each step of the LAR procedure, using a simulated data set with six predictors. The labels at the top of the plot indicate which variables enter the active set at each step. The step length are measured in units of L_1 arc length.*

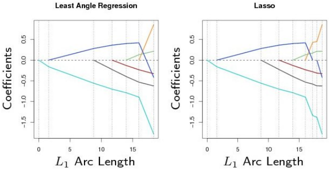


FIGURE 3.15. *Left panel shows the LAR coefficient profiles on the simulated data, as a function of the L_1 arc length. The right panel shows the Lasso profile. They are identical until the dark-blue coefficient crosses zero at an arc length of about 18.*

LARS: Solving LAR with Lasso

- Experimental observation

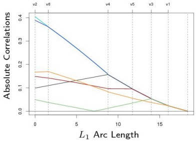


FIGURE 3.14. *Progression of the absolute correlations during each step of the LAR procedure, using a simulated data set with six predictors. The labels at the top of the plot indicate which variables enter the active set at each step. The step length are measured in units of L_1 arc length.*

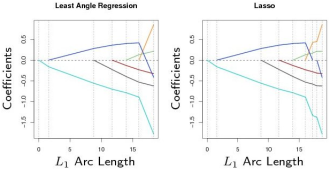


FIGURE 3.15. *Left panel shows the LAR coefficient profiles on the simulated data, as a function of the L_1 arc length. The right panel shows the Lasso profile. They are identical until the dark-blue coefficient crosses zero at an arc length of about 18.*

- So $|\beta_j|$ can begin to decrease, and it can change sign.

LARS: Solving LAR with Lasso

- Experimental observation

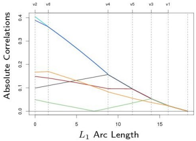


FIGURE 3.14. *Progression of the absolute correlations during each step of the LAR procedure, using a simulated data set with six predictors. The labels at the top of the plot indicate which variables enter the active set at each step. The step length are measured in units of L_1 arc length.*

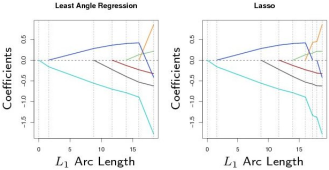


FIGURE 3.15. *Left panel shows the LAR coefficient profiles on the simulated data, as a function of the L_1 arc length. The right panel shows the Lasso profile. They are identical until the dark-blue coefficient crosses zero at an arc length of about 18.*

- So $|\beta_j|$ can begin to decrease, and it can change sign.
- The LARS fix: Drop coefficients which hit 0 out of 'active set'.

Lasso solutions: Observations

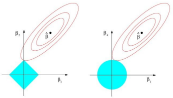


FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions* $|\beta_1| + |\beta_2| \le t$ *and* $\beta_1^2 + \beta_2^2 \le t^2$*, respectively, while the red ellipses are the contours of the least squares error function.*

Lasso solutions: Observations

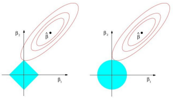


FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions $|\beta_1| + |\beta_2| \leq t$ and $\beta_1^2 + \beta_2^2 \leq t^2$, respectively, while the red ellipses are the contours of the least squares error function.*

- Note: $f(\hat{\beta}) = \min_\beta \|y - X\beta\|_2^2 + \lambda \sum_i |\beta_i|$.

Lasso solutions: Observations

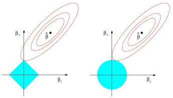


FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions $|\beta_1| + |\beta_2| \le t$ and $\beta_1^2 + \beta_2^2 \le t^2$, respectively, while the red ellipses are the contours of the least squares error function.*

- Note: $f(\hat{\beta}) = \min_\beta \|y - X\beta\|_2^2 + \lambda \sum_i |\beta_i|$.
- Set $\nabla f(\beta) = 0$.

Lasso solutions: Observations

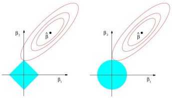


FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions $|\beta_1| + |\beta_2| \le t$ and $\beta_1^2 + \beta_2^2 \le t^2$, respectively, while the red ellipses are the contours of the least squares error function.*

- Note: $f(\hat{\beta}) = \min_\beta \|y - X\beta\|_2^2 + \lambda \sum_i |\beta_i|$.
- Set $\nabla f(\beta) = 0$.
- B := features in sparse solution.

Lasso solutions: Observations

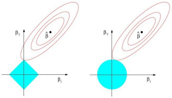


FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions $|\beta_1| + |\beta_2| \leq t$ and $\beta_1^2 + \beta_2^2 \leq t^2$, respectively, while the red ellipses are the contours of the least squares error function.*

- Note: $f(\hat{\beta}) = \min_\beta \|y - X\beta\|_2^2 + \lambda \sum_i |\beta_i|$.
- Set $\nabla f(\beta) = 0$.
- B := features in sparse solution.
- Get conditions:
  $\forall j \in B : x_j^T(y - X\beta) = \lambda * sgn(\beta_j)$.
  $\forall j \notin B : x_j^T(y - X\beta) = \lambda * sgn(\beta_j) \leq |\lambda|$.

Lasso solutions: Observations

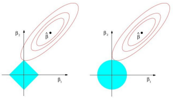


FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions $|\beta_1| + |\beta_2| \leq t$ and $\beta_1^2 + \beta_2^2 \leq t^2$, respectively, while the red ellipses are the contours of the least squares error function.*

- Note: $f(\hat{\beta}) = \min_\beta \|y - X\beta\|_2^2 + \lambda \sum_i |\beta_i|$.
- Set $\nabla f(\beta) = 0$.
- B := features in sparse solution.
- Get conditions:
  $\forall j \in B : x_j^T(y - X\beta) = \lambda * sgn(\beta_j)$.
  $\forall j \notin B : x_j^T(y - X\beta) = \lambda * sgn(\beta_j) \leq |\lambda|$.
- Remarkable! $\lambda \rightarrow$ upper bound on correlation of the residue with x_j.

Compare with LAR

Compare with LAR

- $\hat{c}$: correlation of residue with features. s_j: sign of correlation with feature j.

Compare with LAR

- $\hat{c}$: correlation of residue with features. s_j: sign of correlation with feature j.
- $\forall j \in A : x_j^T(y - X\beta) = \hat{c} * s_j$.

Compare with LAR

- $\hat{c}$: correlation of residue with features. s_j: sign of correlation with feature j.
- $\forall j \in A : x_j^T(y - X\beta) = \hat{c} * s_j$.
- $\forall j \notin A : x_j^T(y - X\beta) \leq \hat{c}$.

Compare with LAR

- $\hat{c}$: correlation of residue with features. s_j: sign of correlation with feature j.
- $\forall j \in A : x_j^T(y - X\beta) = \hat{c} * s_j$.
- $\forall j \notin A : x_j^T(y - X\beta) \leq \hat{c}$.
- Compare:
  $\forall j \in B : x_j^T(y - X\beta) = \lambda * sgn(\beta_j)$.
  $\forall j \notin B : x_j^T(y - X\beta) \leq |\lambda|$.

Compare with LAR

- $\hat{c}$: correlation of residue with features. s_j: sign of correlation with feature j.
- $\forall j \in A : x_j^T(y - X\beta) = \hat{c} * s_j$.
- $\forall j \notin A : x_j^T(y - X\beta) \leq \hat{c}$.
- Compare:
  $\forall j \in B : x_j^T(y - X\beta) = \lambda * sgn(\beta_j)$.
  $\forall j \notin B : x_j^T(y - X\beta) \leq |\lambda|$.
- When $sgn(\beta_j) \neq s_j$, Lasso $\neq$ LAR.

Compare with LAR

- $\hat{c}$: correlation of residue with features. s_j: sign of correlation with feature j.
- $\forall j \in A : x_j^T(y - X\beta) = \hat{c} * s_j$.
- $\forall j \notin A : x_j^T(y - X\beta) \leq \hat{c}$.
- Compare:
  $\forall j \in B : x_j^T(y - X\beta) = \lambda * sgn(\beta_j)$.
  $\forall j \notin B : x_j^T(y - X\beta) \leq |\lambda|$.
- When $sgn(\beta_j) \neq s_j$, Lasso $\neq$ LAR.
- The LARS fix: If β_j hits 0 for $j \in A$, it is about to change sign.

When $sgn(\beta_j) \neq s_j$, Lasso $\neq$ LAR

When $sgn(\beta_j) \neq s_j$, Lasso $\neq$ LAR

- Experimental observation

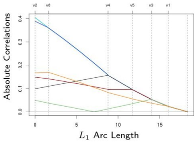


FIGURE 3.14. *Progression of the absolute correlations during each step of the LAR procedure, using a simulated data set with six predictors. The labels at the top of the plot indicate which variables enter the active set at each step. The step length are measured in units of L_1 arc length.*

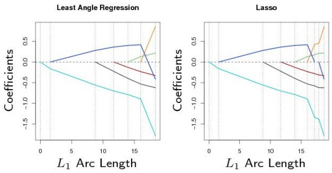


FIGURE 3.15. *Left panel shows the LAR coefficient profiles on the simulated data, as a function of the L_1 arc length. The right panel shows the Lasso profile. They are identical until the dark-blue coefficient crosses zero at an arc length of about 18.*

When $sgn(\beta_j) \neq s_j$, Lasso $\neq$ LAR

- Experimental observation

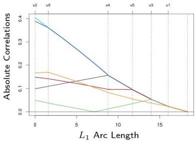


FIGURE 3.14. *Progression of the absolute correlations during each step of the LAR procedure, using a simulated data set with six predictors. The labels at the top of the plot indicate which variables enter the active set at each step. The step length are measured in units of L_1 arc length.*

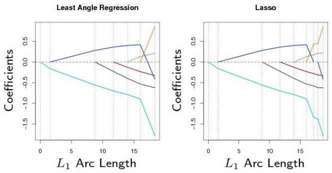


FIGURE 3.15. *Left panel shows the LAR coefficient profiles on the simulated data, as a function of the L_1 arc length. The right panel shows the Lasso profile. They are identical until the dark-blue coefficient crosses zero at an arc length of about 18.*

- Geometric intuition: When does a coefficient start decreasing, even when correlation is positive? When you add feature v1 which is not independent of A. As $\beta_1 \uparrow, \beta_6 \downarrow$.

Outline

1. Outline
2. Least Angle regression
3. LAR and Lasso: the connection
   - Lasso
   - The connection
4. Conclusion

What did we learn?

What did we learn?

- Least Angle regression (LAR): get sparse solutions, like forward stepwise regression; but modified to be non greedy.

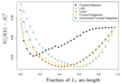


FIGURE 3.16. *Comparison of LAR and lasso with forward stepwise, forward stagewise (FS) and incremental forward stagewise (FS_0) regression. The setup*

What did we learn?

- Least Angle regression (LAR): get sparse solutions, like forward stepwise regression; but modified to be non greedy.

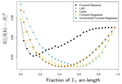


FIGURE 3.16. *Comparison of LAR and lasso with forward stepwise, forward stagewise (FS) and incremental forward stagewise (FS_0) regression. The setup*

- Lasso for sparse solutions.

What did we learn?

- Least Angle regression (LAR): get sparse solutions, like forward stepwise regression; but modified to be non greedy.

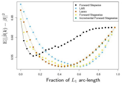


FIGURE 3.16. *Comparison of LAR and lasso with forward stepwise, forward stagewise (FS) and incremental forward stagewise (FS_0) regression. The setup*

- Lasso for sparse solutions.
- The sparse solutions are sometimes different.

What did we learn?

- Least Angle regression (LAR): get sparse solutions, like forward stepwise regression; but modified to be non greedy.

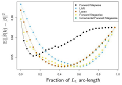


FIGURE 3.16. *Comparison of LAR and lasso with forward stepwise, forward stagewise (FS) and incremental forward stagewise (FS_0) regression. The setup*

- Lasso for sparse solutions.
- The sparse solutions are sometimes different.
- LARS: LAR modified to solve Lasso. Very efficient! Also looked at how Lasso works with new eyes ★★.

References

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani.
Least angle regression.
Annals of Statistics, 32(2):407–499, 2004.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman.
The Elements of Statistical Learning, Second Edition: Data Mining, Inference, and Prediction.
Springer Series in Statistics. Springer, 2nd ed. 2009. corr. 3rd printing edition, September 2009.

Bye!

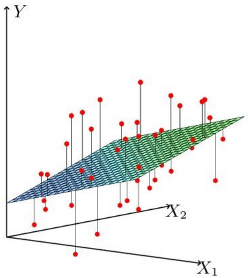


FIGURE 3.1. *Linear least squares fitting with $X \in \mathbb{R}^2$. We seek the linear function of X that minimizes the sum of squared residuals from Y.*

Bye!

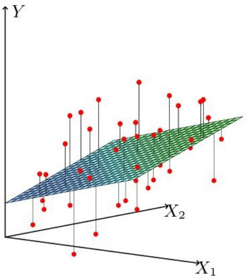


FIGURE 3.1. *Linear least squares fitting with $X \in \mathbb{R}^2$. We seek the linear function of X that minimizes the sum of squared residuals from Y.*

Ask us some questions!