On the Distribution of the Adaptive LASSO Estimator – part II

Ulrike Schneider

University of Vienna

GK, University of Göttingen
January 15, 2009
Outline

1. More on penalized LS (ML) estimators.
2. Penalized LS with orthogonal design.
4. Computational issues.
Penalized LS (ML) estimators

Linear regression model

\[ y = \theta_1 x_1 + \ldots + \theta_k x_k + \varepsilon \]

- response \( y \in \mathbb{R}^n \)
- regressors \( x_i \in \mathbb{R}^n, 1 \leq i \leq k \)
- errors \( \varepsilon \in \mathbb{R}^n \)
- (unknown) parameter vector \( \theta = (\theta_1, \ldots, \theta_k)' \in \mathbb{R}^k \)

A penalized least-squares (LS) estimator \( \hat{\theta} \) for \( \theta \) is given by

\[
\hat{\theta} = \arg \min_{\theta \in \mathbb{R}^k} \left( \frac{1}{2} \| y - X\theta \|^2 + p(\theta) \right)
\]

The penalty function \( p(\theta) \) involves a tuning parameter \( \lambda_n \) (\( \lambda_n = 0 \) corresponds to unpenalized/ordinary LS). \( X = [x_1, \ldots, x_k] \) the \( n \times k \) regression matrix.
Penalized LS (ML) estimators

Linear regression model

$$y = \theta_1 x_1 + \ldots \theta_k x_k + \varepsilon$$

- response $y \in \mathbb{R}^n$
- regressors $x_i \in \mathbb{R}^n$, $1 \leq i \leq k$
- errors $\varepsilon \in \mathbb{R}^n$
- (unknown) parameter vector $\theta = (\theta_1, \ldots, \theta_k)' \in \mathbb{R}^k$

A penalized least-squares (LS) estimator $\hat{\theta}$ for $\theta$ is given by

$$\hat{\theta} = \arg \min_{\theta \in \mathbb{R}^k} \|y - X\theta\|^2 + p(\theta)$$

The penalty function $p(\theta)$ involves a tuning parameter $\lambda_n$ ($\lambda_n = 0$ corresponds to unpenalized/ordinary LS).

$X = [x_1, \ldots, x_k]$ the $n \times k$ regression matrix.
Clearly, different penalties give rise to different estimators.

- General class of Bridge-estimators (Frank & Friedman, 1993) using $l_\gamma$ - type penalties

$$p(\theta) = \lambda_n \sum_{i=1}^{k} |\theta_i|^\gamma$$

$\gamma = 2$: Ridge-estimator (Hoerl & Kennard, 1970)

$\gamma = 1$: LASSO (Tibshirani, 1996).

- Hard- and soft-thresholding estimators.

- Smoothly clipped absolute deviation (SCAD) estimator (Fan & Li, 2001).

- Adaptive LASSO estimator (Zou, 2006).
Clearly, different penalties give rise to different estimators.

- General class of Bridge-estimators (Frank & Friedman, 1993) using $l_{\gamma}$-type penalties

$$p(\theta) = \lambda_n \sum_{i=1}^{k} |\theta_i|^\gamma$$

- $\gamma = 2$: Ridge-estimator (Hoerl & Kennard, 1970)
- $\gamma = 1$: LASSO (Tibshirani, 1996).

- Hard- and soft-thresholding estimators.

- Smoothly clipped absolute deviation (SCAD) estimator (Fan & Li, 2001).

- Adaptive LASSO estimator (Zou, 2006).
Clearly, different penalties give rise to different estimators.

- General class of Bridge-estimators (Frank & Friedman, 1993) using $l_\gamma$ - type penalties
  \[
p(\theta) = \lambda_n \sum_{i=1}^{k} |\theta_i|^\gamma
  \]
  
  $\gamma = 2$: Ridge-estimator (Hoerl & Kennard, 1970)
  
  $\gamma = 1$: LASSO (Tibshirani, 1996).

- Hard- and soft-thresholding estimators.
  - Smoothly clipped absolute deviation (SCAD) estimator (Fan & Li, 2001).
  - Adaptive LASSO estimator (Zou, 2006).
Clearly, different penalties give rise to different estimators.

- General class of Bridge-estimators (Frank & Friedman, 1993) using $l_\gamma$-type penalties
  
  $$p(\theta) = \lambda_n \sum_{i=1}^{k} |\theta_i|^\gamma$$

  $\gamma = 2$: Ridge-estimator (Hoerl & Kennard, 1970)
  $\gamma = 1$: LASSO (Tibshirani, 1996).

- Hard- and soft-thresholding estimators.

- Smoothly clipped absolute deviation (SCAD) estimator (Fan & Li, 2001).

- Adaptive LASSO estimator (Zou, 2006).
Clearly, different penalties give rise to different estimators.

- General class of Bridge-estimators (Frank & Friedman, 1993) using $l_\gamma$ - type penalties

$$ p(\theta) = \lambda_n \sum_{i=1}^{k} |\theta_i|^\gamma $$

$\gamma = 2$: Ridge-estimator (Hoerl & Kennard, 1970)

$\gamma = 1$: LASSO (Tibshirani, 1996).

- Hard- and soft-thresholding estimators.

- Smoothly clipped absolute deviation (SCAD) estimator (Fan & Li, 2001).

- Adaptive LASSO estimator (Zou, 2006).
Relationship to classical PMS estimators

Brigde-estimators satisfy

$$\min \|y - X\theta\|^2 + \lambda_n \sum_{i=1}^k |\theta_i|\gamma \quad (0 < \gamma < \infty)$$

For $\gamma \to 0$, get

$$\min \|y - X\theta\|^2 + \lambda_n \text{card}\{i : \theta_i \neq 0\}$$

which yields a minimum $C_p$-type procedure such as AIC and BIC. ($l_\gamma$-type penalty with “$\gamma = 0$”)
For “$\gamma = 0$” procedures are computationally expensive.

For $\gamma > 0$ (Bridge) estimators are more computationally tractable, especially for $\gamma \geq 1$ (convex objective function).

For $\gamma \leq 1$, estimators perform model selection

$$P(\hat{\theta}_i = 0) > 0 \text{ if } \theta_i = 0$$

Same for SCAD, hard- and soft-thresholding. Phenomenon is more pronounced for smaller $\gamma$.

$\gamma = 1$ (LASSO and adaptive LASSO) as compromise between the wish to detect zeros and computational simplicity. (SCAD leads to a non-convex optimization problem.)
For “\( \gamma = 0 \)” procedures are computationally expensive.

For \( \gamma > 0 \) (Bridge) estimators are more computationally tractable, especially for \( \gamma \geq 1 \) (convex objective function).

For \( \gamma \leq 1 \), estimators perform model selection

\[
P(\hat{\theta}_i = 0) > 0 \quad \text{if} \quad \theta_i = 0
\]

Same for SCAD, hard- and soft-thresholding. Phenomenon is more pronounced for smaller \( \gamma \).

\( \gamma = 1 \) (LASSO and adaptive LASSO) as compromise between the wish to detect zeros and computational simplicity. (SCAD leads to a non-convex optimization problem.)
For “$\gamma = 0$” procedures are computationally expensive.

For $\gamma > 0$ (Bridge) estimators are more computationally tractable, especially for $\gamma \geq 1$ (convex objective function).

For $\gamma \leq 1$, estimators perform model selection

$$P(\hat{\theta}_i = 0) > 0 \text{ if } \theta_i = 0$$

Same for SCAD, hard- and soft-thresholding. Phenomenon is more pronounced for smaller $\gamma$.

$\gamma = 1$ (LASSO and adaptive LASSO) as compromise between the wish to detect zeros and computational simplicity. (SCAD leads to a non-convex optimization problem.)
For “$\gamma = 0$” procedures are computationally expensive.

For $\gamma > 0$ (Bridge) estimators are more computationally tractable, especially for $\gamma \geq 1$ (convex objective function).

For $\gamma \leq 1$, estimators perform model selection

$$P(\hat{\theta}_i = 0) > 0 \quad \text{if} \quad \theta_i = 0$$

Same for SCAD, hard- and soft-thresholding. Phenomenon is more pronounced for smaller $\gamma$.

$\gamma = 1$ (LASSO and adaptive LASSO) as compromise between the wish to detect zeros and computational simplicity. (SCAD leads to a non-convex optimization problem.)
Orthogonal design

Linear regression model

\[ y = \theta_1 x_1 + \ldots \theta_k x_k + \varepsilon \]

- \( X \) is non-stochastic, \( n \times k \) and \( rk(X) = k \).
- \( \varepsilon \sim N_n(0, \sigma^2 I_n) \)
- \( \sigma^2 \) is known (wlog \( \sigma^2 = 1 \)) and \( X'X \) is diagonal, in particular \( X'X = nI_k \).

Again, wlog consider Gaussian location model \( y_1, \ldots, y_n \overset{iid}{\sim} N(\theta, 1) \).

Then \( \hat{\theta}_{OLS} = \hat{\theta}_{MLE} = \bar{y} \) and we want to choose between the restricted model \( M_R = \{N(0,1)\} \) or the full model \( M_U = \{N(\theta,1) : \theta \in \mathbb{R}\} \).
Hard-thresholding $\hat{\theta}_H$

\[ p(\theta) = n \left[ \mu_n^2 - (|\theta| - \mu_n)^2 \mathbf{1}(|\theta| < \mu_n) \right] \]

\[ \hat{\theta}_H = \bar{y} \mathbf{1}(|\bar{y}| > \mu_n) \]

- Equivalent to a post-model estimator based on (eg) t-tests.
- Estimator is not continuous.
- Possesses an “oracle-property” if sparsely-tuned.
Soft-thresholding $\hat{\theta}_L$

\[ p(\theta) = 2n\mu_n|\theta| \]
\[ \hat{\theta}_L = \text{sign}(\bar{y}) (|\bar{y}| - \mu_n)_+ \]

- Equivalent to LASSO.
- Bias problem! No “oracle-property”.

LASSO
Smoothly-clipped-absolute-deviation $\hat{\theta}_{SCAD}$

$$p'(\theta) = \mu_n \left[ 1(\theta \leq \mu_n) + (a\mu_n - \theta)_+ / ((a-1)\mu_n) 1(\theta > \mu_n) \right],$$

where $a > 2$ is an additional tuning parameter.

$$\hat{\theta}_{SCAD} = \begin{cases} 
\text{sign}(\bar{y})(|\bar{y}| - \mu_n)_+ & \text{if } |\bar{y}| \leq 2\mu_n \\
(a-1)\bar{y} - \text{sign}(\bar{y})a\mu_n / (a-2) & \text{if } 2\mu_n < |\bar{y}| \leq a\mu_n \\
\bar{y} & \text{if } |\bar{y}| > a\mu_n 
\end{cases}$$

- Non-convex optimization problem.
- Possesses an “oracle-property” if sparsely-tuned.
Adaptive LASSO $\hat{\theta}_{AL}$

\[ p(\theta) = 2n\mu_n^2|\theta|/|\bar{y}| \]

\[ \hat{\theta}_{AL} = \begin{cases} 0 & \text{if } |\bar{y}| \leq \mu_n \\ \bar{y} - \mu_n^2/\bar{y} & \text{if } |\bar{y}| > \mu_n \end{cases} \]

- Equivalent to non-negative Garotte (Breiman, 1995)
- Possesses an “oracle-property” if sparsely-tuned.
Why moving-parameter asymptotics?

Let’s you see what’s really going on in large samples if the convergence is not uniform with respect the underlying parameter.

- The unpenalized LS estimator is $\hat{\theta}_{OLS} = \bar{y}$ in our model with $\hat{\theta}_{OLS} \sim N(\theta, 1/n)$, so that

  \[
  n^{1/2}(\hat{\theta}_{OLS} - \theta) \sim N(0, 1)
  \]

  for each sample size $n \in \mathbb{N}$, so the distribution is independent of $\theta$.

- For $\hat{\theta}_{AL}$ (and other PLSEs), the distribution of $n^{1/2}(\hat{\theta}_{AL} - \theta)$ depends on $\theta$ in a complicated manner.
Why moving-parameter asymptotics?

Let’s see what’s really going on in large samples if the convergence is not uniform with respect to the underlying parameter.

- The unpenalized LS estimator is $\hat{\theta}_{\text{OLS}} = \bar{y}$ in our model with $\hat{\theta}_{\text{OLS}} \sim \mathcal{N}(\theta, 1/n)$, so that

$$n^{1/2}(\hat{\theta}_{\text{OLS}} - \theta) \sim \mathcal{N}(0, 1)$$

for each sample size $n \in \mathbb{N}$, so the distribution is independent of $\theta$.

- For $\hat{\theta}_{\text{AL}}$ (and other PLSEs), the distribution of $n^{1/2}(\hat{\theta}_{\text{AL}} - \theta)$ depends on $\theta$ in a complicated manner.
Why moving-parameter asymptotics?

Let’s you see what’s really going on in large samples if the convergence is not uniform with respect the underlying parameter.

- The unpenalized LS estimator is $\hat{\theta}_{\text{OLS}} = \bar{y}$ in our model with $\hat{\theta}_{\text{OLS}} \sim N(\theta, 1/n)$, so that
  \[ n^{1/2}(\hat{\theta}_{\text{OLS}} - \theta) \sim N(0, 1) \]
  for each sample size $n \in \mathbb{N}$, so the distribution is independent of $\theta$.

- For $\hat{\theta}_{\text{AL}}$ (and other PLSEs), the distribution of $n^{1/2}(\hat{\theta}_{\text{AL}} - \theta)$ depends on $\theta$ in a complicated manner.
Why moving-parameter asymptotics?

Let’s you see what’s really going on in large samples if the convergence is not uniform with respect the underlying parameter.

- The unpenalized LS estimator is $\hat{\theta}_{OLS} = \bar{y}$ in our model with $\hat{\theta}_{OLS} \sim N(\theta, 1/n)$, so that
  
  $$n^{1/2}(\hat{\theta}_{OLS} - \theta) \sim N(0, 1)$$

  for each sample size $n \in \mathbb{N}$, so the distribution is independent of $\theta$.

- For $\hat{\theta}_{AL}$ (and other PLSEs), the distribution of $n^{1/2}(\hat{\theta}_{AL} - \theta)$ depends on $\theta$ in a complicated manner.

\[
1(n^{1/2}\theta + x \geq 0) \Phi \left(\frac{-(n^{1/2}\theta - x)}{2} + \sqrt{((n^{1/2}\theta + x)/2)^2 + n\mu_n^2}\right) + 1(n^{1/2}\theta + x < 0) \Phi \left(\frac{-(n^{1/2}\theta - x)}{2} - \sqrt{((n^{1/2}\theta + x)/2)^2 + n\mu_n^2}\right)
\]
Why moving-parameter asymptotics?

Let’s see what’s really going on in large samples if the convergence is not uniform with respect the underlying parameter.

- The unpenalized LS estimator is $\hat{\theta}_{\text{OLS}} = \bar{y}$ in our model with $\hat{\theta}_{\text{OLS}} \sim N(\theta, 1/n)$, so that

  $$n^{1/2}(\hat{\theta}_{\text{OLS}} - \theta) \sim N(0, 1)$$

  for each sample size $n \in \mathbb{N}$, so the distribution is independent of $\theta$.

- For $\hat{\theta}_{\text{AL}}$ (and other PLSEs), the distribution of $n^{1/2}(\hat{\theta}_{\text{AL}} - \theta)$ depends on $\theta$ in a complicated manner.

- Even for large $n$, the pointwise asymptotic distribution might be “far” from the finite-sample distribution of interest if the underlying convergence is not uniform, as we have seen yesterday.
Asymptotic model selection probabilities

Probability of choosing the restricted model $M_R$ is given by

$$P_{n,\theta}(\hat{\theta} = 0) = \Phi\left(-\frac{n^{1/2}}{2}(\theta + \mu_n)\right) - \Phi\left(-\frac{n^{1/2}}{2}(\theta - \mu_n)\right),$$

and clearly, the probability of choosing the unrestricted model $M_U$ is

$$P_{n,\theta}(\hat{\theta} \neq 0) = 1 - P_{n,\theta}(\hat{\theta} = 0)$$

($\hat{\theta}$ any of the previous PLS estimators).
Asymptotic model selection probabilities

\[ n = 1, \quad \mu_n = n^{-1/3} \text{ (consistent case)} \]
Asymptotic model selection probabilities

\[ n = 2, \quad \mu_n = n^{-1/3} \quad \text{(consistent case)} \]
Asymptotic model selection probabilities

\[ n = 3, \quad \mu_n = n^{-1/3} \text{ (consistent case)} \]
Asymptotic model selection probabilities

\[ n = 4, \quad \mu_n = n^{-1/3} \text{ (consistent case)} \]
Asymptotic model selection probabilities

\[ n = 5, \quad \mu_n = n^{-1/3} \text{ (consistent case)} \]
Asymptotic model selection probabilities

\[ n = 7, \quad \mu_n = n^{-1/3} \text{ (consistent case)} \]
Asymptotic model selection probabilities

\( n = 10, \quad \mu_n = n^{-1/3} \) (consistent case)
Asymptotic model selection probabilities

\[ n = 20, \quad \mu_n = n^{-1/3} \text{ (consistent case)} \]
Asymptotic model selection probabilities

\[ n = 50, \quad \mu_n = n^{-1/3} \text{ (consistent case)} \]
Asymptotic model selection probabilities

\[ n = 70, \quad \mu_n = n^{-1/3} \text{ (consistent case)} \]
Asymptotic model selection probabilities

\[ n = 100, \quad \mu_n = n^{-1/3} \text{ (consistent case)} \]
Asymptotic model selection probabilities

\[ n = 200, \quad \mu_n = n^{-1/3} \text{ (consistent case)} \]
Asymptotic model selection probabilities

\[ n = 500, \quad \mu_n = n^{1/3} \text{ (consistent case)} \]
Asymptotic model selection probabilities

\[ n = 1000, \quad \mu_n = n^{-1/3} \text{ (consistent case)} \]
Asymptotic model selection probabilities

\[ n = 2000, \quad \mu_n = n^{-1/3} \text{ (consistent case)} \]
Asymptotic model selection probabilities

\[ n = 5000, \quad \mu_n = n^{-1/3} \text{ (consistent case)} \]
Asymptotic model selection probabilities

\[ n = 10000, \mu_n = n^{-1/3} \] (consistent case)
Asymptotic model selection probabilities

\[ n = \infty, \quad \mu_n = n^{-1/3} \text{ (consistent case)} \]
Model selection probabilities

1. **Consistent case** \((\mu_n \to 0, n^{1/2}\mu_n \to \infty)\)

Assume \(\theta_n/\mu_n \to \zeta \in \mathbb{R} \cup \{-\infty, \infty\}\). Then

\[
\lim_{n \to \infty} P_{n, \theta_n}(\hat{\theta}_{AL} = 0) =
\begin{cases} 
1 & \text{if } |\zeta| < 1 \\
\Phi(r) & \text{if } |\zeta| = 1, n^{1/2}(\mu_n - \zeta \theta_n) \to r \in \mathbb{R} \cup \{-\infty, \infty\} \\
0 & \text{if } |\zeta| > 1
\end{cases}
\]

Deviations of \(\theta_n\) from 0 of order \(n^{-1/2}\) are not detected at all!

2. **Conservative case** \((\mu_n \to 0, n^{1/2}\mu_n \to m, 0 \leq m < \infty)\)

Assume \(\theta_n \in \mathbb{R}\) satisfies \(n^{1/2}\theta_n \to \nu \in \mathbb{R} \cup \{-\infty, \infty\}\). Then

\[
\lim_{n \to \infty} P_{n, \theta_n}(\hat{\theta}_{AL} = 0) = \Phi(-\nu + m) - \Phi(-\nu - m).
\]

Deviations of \(\theta_n\) from 0 of order \(n^{-1/2}\) are detected with positive prob.
Model selection probabilities

1. **Consistent case** \((\mu_n \to 0, \ n^{1/2}\mu_n \to \infty)\)

Assume \(\theta_n/\mu_n \to \zeta \in \mathbb{R} \cup \{-\infty, \infty\}\). Then

\[
\lim_{n \to \infty} P_{n,\theta_n}(\hat{\theta}_{AL} = 0) =
\begin{cases}
1 & \text{if } |\zeta| < 1 \\
\Phi(r) & \text{if } |\zeta| = 1, \ n^{1/2}(\mu_n - \zeta \theta_n) \to r \in \mathbb{R} \cup \{-\infty, \infty\} \\
0 & \text{if } |\zeta| > 1
\end{cases}
\]

Deviations of \(\theta_n\) from 0 of order \(n^{-1/2}\) are not detected at all!

2. **Conservative case** \((\mu_n \to 0, \ n^{1/2}\mu_n \to m, \ 0 \leq m < \infty)\)

Assume \(\theta_n \in \mathbb{R}\) satisfies \(n^{1/2}\theta_n \to \nu \in \mathbb{R} \cup \{-\infty, \infty\}\). Then

\[
\lim_{n \to \infty} P_{n,\theta_n}(\hat{\theta}_{AL} = 0) = \Phi(-\nu + m) - \Phi(-\nu - m).
\]

Deviations of \(\theta_n\) from 0 of order \(n^{-1/2}\) are detected with positive prob.
Model selection probabilities

1. **Consistent case** \((\mu_n \to 0, \, n^{1/2}\mu_n \to \infty)\)

Assume \(\theta_n/\mu_n \to \zeta \in \mathbb{R} \cup \{-\infty, \infty\}\). Then

\[
\lim_{n \to \infty} P_{n,\theta_n}(\hat{\Theta}_{AL} = 0) = \\
\begin{cases} 
1 & \text{if } |\zeta| < 1 \\
\Phi(r) & \text{if } |\zeta| = 1, \, n^{1/2}(\mu_n - \zeta\theta_n) \to r \in \mathbb{R} \cup \{-\infty, \infty\} \\
0 & \text{if } |\zeta| > 1 
\end{cases}
\]

Deviations of \(\theta_n\) from 0 of order \(n^{-1/2}\) are not detected at all!

2. **Conservative case** \((\mu_n \to 0, \, n^{1/2}\mu_n \to m, \, 0 \leq m < \infty)\)

Assume \(\theta_n \in \mathbb{R}\) satisfies \(n^{1/2}\theta_n \to \nu \in \mathbb{R} \cup \{-\infty, \infty\}\). Then

\[
\lim_{n \to \infty} P_{n,\theta_n}(\hat{\Theta}_{AL} = 0) = \Phi(-\nu + m) - \Phi(-\nu - m). 
\]

Deviations of \(\theta_n\) from 0 of order \(n^{-1/2}\) are detected with positive prob.
Model selection probabilities

1. **Consistent case** \( (\mu_n \to 0, n^{1/2}\mu_n \to \infty) \)

   Assume \( \theta_n/\mu_n \to \zeta \in \mathbb{R} \cup \{-\infty, \infty\} \). Then

   \[
   \lim_{n \to \infty} P_{n,\theta_n}(\hat{\theta}_{AL} = 0) =
   \begin{cases}
   1 & \text{if } |\zeta| < 1 \\
   \Phi(r) & \text{if } |\zeta| = 1, n^{1/2}(\mu_n - \zeta \theta_n) \to r \in \mathbb{R} \cup \{-\infty, \infty\} \\
   0 & \text{if } |\zeta| > 1
   \end{cases}
   \]

   Deviations of \( \theta_n \) from 0 of order \( n^{-1/2} \) are not detected at all!

2. **Conservative case** \( (\mu_n \to 0, n^{1/2}\mu_n \to m, 0 \leq m < \infty) \)

   Assume \( \theta_n \in \mathbb{R} \) satisfies \( n^{1/2}\theta_n \to \nu \in \mathbb{R} \cup \{-\infty, \infty\} \). Then

   \[
   \lim_{n \to \infty} P_{n,\theta_n}(\hat{\theta}_{AL} = 0) = \Phi(-\nu + m) - \Phi(-\nu - m).
   \]

   Deviations of \( \theta_n \) from 0 of order \( n^{-1/2} \) are detected with positive prob.
Consistent procedures cannot uncover deviations from zero of order $n^{-1/2}$. This matters e.g. since usually $n^{1/2}(\hat{\theta} - \theta)$ is considered.

Conservative procedures do detect such deviations with positive probability.

Often the parameter space is assumed to be bounded away from zero by a rate smaller than $n^{-1/2}$.

Model selection is “hard” when the true parameter $\theta$ is close to zero! (Yet this is an interesting case.)
Consistent procedures cannot uncover deviations from zero of order $n^{-1/2}$. This matters e.g. since usually $n^{1/2}(\hat{\theta} - \theta)$ is considered.

Conservative procedures do detect such deviations with positive probability.

Often the parameter space is assumed to be bounded away from zero by a rate smaller than $n^{-1/2}$.

Model selection is “hard” when the true parameter $\theta$ is close to zero! (Yet this is an interesting case.)
Consistent procedures cannot uncover deviations from zero of order $n^{-1/2}$. This matters e.g. since usually $n^{1/2}(\hat{\theta} - \theta)$ is considered.

Conservative procedures do detect such deviations with positive probability.

Often the parameter space is assumed to be bounded away from zero by a rate smaller than $n^{-1/2}$.

Model selection is “hard” when the true parameter $\theta$ is close to zero! (Yet this is an interesting case.)
Consistent procedures cannot uncover deviations from zero of order $n^{-1/2}$. This matters e.g. since usually $n^{1/2}(\hat{\theta} - \theta)$ is considered.

Conservative procedures do detect such deviations with positive probability.

Often the parameter space is assumed to be bounded away from zero by a rate smaller than $n^{-1/2}$.

Model selection is “hard” when the true parameter $\theta$ is close to zero! (Yet this is an interesting case.)
Why does LASSO perform model selection?

Rewrite minimization problem $\min_{\theta \in \mathbb{R}^k} \|y - X\theta\|^2 + \lambda_n \sum_{i=1}^{k} |\theta_i|$ as
Why does LASSO perform model selection?

Rewrite minimization problem \( \min_{\theta \in \mathbb{R}^k} \| y - X \theta \|^2 + \lambda_n \sum_{i=1}^{k} |\theta_i| \) as

\[
\begin{align*}
\min_{\theta \in \mathbb{R}^k} & \quad \| y - X \theta \|^2 \\
\text{s.t.} & \quad \sum_{i=1}^{k} |\theta_i| \leq s \quad (\text{for some } s \geq 0)
\end{align*}
\]
Why does LASSO perform model selection?

Rewrite minimization problem $\min_{\theta \in \mathbb{R}^k} ||y - X\theta||^2 + \lambda_n \sum_{i=1}^{k} |\theta_i|$ as

$$\min_{\theta \in \mathbb{R}^k} ||y - X\theta||^2$$

s.t. $\sum_{i=1}^{k} |\theta_i| \leq s$ (for some $s \geq 0$)

(Plot from Tibshirani (1996))
Computational issues for (adaptive) LASSO

- Clearly, the LASSO estimator $\hat{\theta}_L$ depends on the tuning parameter $\lambda_n$.
- The “solution paths” for each component $\hat{\theta}_{L,i}(\lambda_n)$ can be shown to be piecewise linear in $\lambda_n$ for each $i = 1, \ldots, k$. (Rosset and Zhu, 2007)
- This property can be exploited to derive efficient algorithms to compute $\hat{\theta}_L$ “easily” for all $\lambda_n \geq 0$ “at once”.
- There exist R-packages to do this, such as the lars package by Efron et al. (2004).
- The adaptive LASSO can be computed from the LASSO solutions using an appropriately transformed regression matrix $X^*$.
Clearly, the LASSO estimator $\hat{\theta}_L$ depends on the tuning parameter $\lambda_n$.

The “solution paths” for each component $\hat{\theta}_{L,i}(\lambda_n)$ can be shown to be piecewise linear in $\lambda_n$ for each $i = 1, \ldots, k$. (Rosset and Zhu, 2007)

This property can be exploited to derive efficient algorithms to compute $\hat{\theta}_L$ “easily” for all $\lambda_n \geq 0$ “at once”.

There exist R-packages to do this, such as the lars package by Efron et al. (2004).

The adaptive LASSO can be computed from the LASSO solutions using an appropriately transformed regression matrix $X^*$.
Clearly, the LASSO estimator $\hat{\theta}_L$ depends on the tuning parameter $\lambda_n$.

The “solution paths” for each component $\hat{\theta}_{L,i}(\lambda_n)$ can be shown to be piecewise linear in $\lambda_n$ for each $i = 1, \ldots, k$. (Rosset and Zhu, 2007)

This property can be exploited to derive efficient algorithms to compute $\hat{\theta}_L$ “easily” for all $\lambda_n \geq 0$ “at once”.

There exist R-packages to do this, such as the lars package by Efron et al. (2004).

The adaptive LASSO can be computed from the LASSO solutions using an appropriately transformed regression matrix $X^*$.
Clearly, the LASSO estimator $\hat{\theta}_L$ depends on the tuning parameter $\lambda_n$.

The “solution paths” for each component $\hat{\theta}_{L,i}(\lambda_n)$ can be shown to be piecewise linear in $\lambda_n$ for each $i = 1, \ldots, k$. (Rosset and Zhu, 2007)

This property can be exploited to derive efficient algorithms to compute $\hat{\theta}_L$ “easily” for all $\lambda_n \geq 0$ “at once”.

There exist R-packages to do this, such as the lars package by Efron et al. (2004).

The adaptive LASSO can be computed from the LASSO solutions using an appropriately transformed regression matrix $X^*$.
Clearly, the LASSO estimator $\hat{\theta}_L$ depends on the tuning parameter $\lambda_n$.

The “solution paths” for each component $\hat{\theta}_{L,i}(\lambda_n)$ can be shown to be piecewise linear in $\lambda_n$ for each $i = 1, \ldots, k$. (Rosset and Zhu, 2007)

This property can be exploited to derive efficient algorithms to compute $\hat{\theta}_L$ “easily” for all $\lambda_n \geq 0$ “at once”.

There exist R-packages to do this, such as the lars package by Efron et al. (2004).

The adaptive LASSO can be computed from the LASSO solutions using an appropriately transformed regression matrix $X^*$.
Clearly, the LASSO estimator $\hat{\theta}_L$ depends on the tuning parameter $\lambda_n$.

The “solution paths” for each component $\hat{\theta}_{L,i}(\lambda_n)$ can be shown to be piecewise linear in $\lambda_n$ for each $i = 1, \ldots, k$. (Rosset and Zhu, 2007)

This property can be exploited to derive efficient algorithms to compute $\hat{\theta}_L$ “easily” for all $\lambda_n \geq 0$ “at once”.

There exist R-packages to do this, such as the lars package by Efron et al. (2004).

The adaptive LASSO can be computed from the LASSO solutions using an appropriately transformed regression matrix $X^*$.
Choosing the tuning parameter

$\lambda_n$ is usually chosen after computing the solutions paths $\hat{\theta}_L(\lambda_n)$, most often by

- generalized cross-validation (minimizing prediction error) generally leads to conservative model selection ① or by using a

- BIC-type criterion (after LASSO) leads to consistent model selection ②
Choosing the tuning parameter

$\lambda_n$ is usually chosen after computing the solutions paths $\hat{\theta}_L(\lambda_n)$, most often by

- **generalized cross-validation** (minimizing prediction error) generally leads to conservative model selection \(^1\) or by using a

- **BIC-type criterion** (after LASSO) leads to consistent model selection \(^2\)
Choosing the tuning parameter

$\lambda_n$ is usually chosen after computing the solutions paths $\hat{\theta}_L(\lambda_n)$, most often by

- **generalized cross-validation** (minimizing prediction error) generally leads to conservative model selection ¹ or by using a

- **BIC-type criterion** (after LASSO) leads to consistent model selection ²
Choosing the tuning parameter
Summary

- Reviewed at PLS estimators and their connection to classical PMS estimators. Some PLSEs coincide with certain PMS estimators in a normal orthogonal linear regression model.

- Discussed moving-parameter framework and that it is needed if convergence is not uniform with respect to the underlying parameter.

- Presented results for model selection probabilities of PLEs. Model selection is “difficult” when the true parameter is close to zero. Conservative procedures “work better” than consistent ones in detecting small parameters to be not equal to zero.

- Looked at computational issues for the (adaptive) LASSO.
Reviewed at PLS estimators and their connection to classical PMS estimators. Some PLSEs coincide with certain PMS estimators in a normal orthogonal linear regression model.

Discussed moving-parameter framework and that it is needed if convergence is not uniform with respect to the underlying parameter.

Presented results for model selection probabilities of PLEs. Model selection is “difficult” when the true parameter is close to zero. Conservative procedures “work better” than consistent ones in detecting small parameters to be not equal to zero.

Looked at computational issues for the (adaptive) LASSO.
Reviewed at PLS estimators and their connection to classical PMS estimators. Some PLSEs coincide with certain PMS estimators in a normal orthogonal linear regression model.

Discussed moving-parameter framework and that it is needed if convergence is not uniform with respect to the underlying parameter.

Presented results for model selection probabilities of PLEs. Model selection is “difficult” when the true parameter is close to zero. Conservative procedures “work better” than consistent ones in detecting small parameters to be not equal to zero.

Looked at computational issues for the (adaptive) LASSO.
Reviewed at PLS estimators and their connection to classical PMS estimators. Some PLSEs coincide with certain PMS estimators in a normal orthogonal linear regression model.

Discussed moving-parameter framework and that it is needed if convergence is not uniform with respect to the underlying parameter.

Presented results for model selection probabilities of PLEs. Model selection is “difficult” when the true parameter is close to zero. Conservative procedures “work better” than consistent ones in detecting small parameters to be not equal to zero.

Looked at computational issues for the (adaptive) LASSO.
References


