On the Distribution of the Adaptive LASSO Estimator (pt I)

Ulrike Schneider
(joint with B.M. Pötscher)

University of Vienna

Stochastics Colloquium, University of Göttingen
January 14, 2009

Outline

Penalized LS (ML) Estimators

Linear regression model

$$\mathbf{y} = \theta_1 \mathbf{x}_{.1} + \ldots \theta_k \mathbf{x}_{.k} + \boldsymbol{\varepsilon}$$

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

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

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^k} \underbrace{\|\mathbf{y} - X\boldsymbol{\theta}\|^2}_{\text{likelihood or LS -part}} + \underbrace{\lambda_n \, p(\boldsymbol{\theta})}_{\text{penalty}}$$

$\lambda_n > 0$ is a tuning parameter ($\lambda_n = 0$ corresponds to unpenalized/ ordinary LS), $X = [\mathbf{x}_{.1}, \ldots, \mathbf{x}_{.k}]$ the $n \times k$ regression matrix.

Penalized LS (ML) Estimators (cont'd)

Clearly, different penalties give rise to different estimators.

- General class of Bridge-estimators (Frank & Friedman, 1993) using l_γ - type penalties

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

$\gamma = 2$: 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

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

For $\gamma \to 0$, get

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^k} \|y - X\boldsymbol{\theta}\|^2 + \lambda_n \operatorname{card}\{i : \theta_i \neq 0\}$$

which yields a minimum C_p-type procedure such as AIC and BIC.
(l_γ-type penalty with "$\gamma = 0$")

Relationship to classical PMS-estimators (cont'd)

- 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

  $$P_{n,\theta}(\hat{\theta}_i = 0) > 0 \quad \text{if } \theta_i = 0.$$

  Same for SCAD, hard- and soft-thresholding. Phenomenon is more pronounced for smaller γ.
- $\gamma = 1$ (LASSO and adaptive LASSO) as compromise between the wish to detect zeros and computational simplicity. (SCAD leads to a non-convex optimization problem.)

The PLS estimator(s) we treat in the following can be viewed to simultaneously perform model selection and parameter estimation.

Some terminology (model selection)

- Consistent model selection – Zero coefficients are found with asymptotic probability equal to 1.

$$\lim_{n\to\infty} P_{n,\theta}(\hat{\theta}_i = 0) = 1 \quad \text{whenever } \theta_i = 0 \quad (1 \le i \le k)$$

$$\lim_{n\to\infty} P_{n,\theta}(\hat{\theta}_i = 0) = 0 \quad \text{whenever } \theta_i \ne 0 \quad (1 \le i \le k)$$

  An estimator performing consistent model selection is said to have the sparsity property.

- Conservative model selection – Zero coefficients are found with asymptotic probability less than 1.

$$\lim_{n\to\infty} P_{n,\theta}(\hat{\theta}_i = 0) < 1 \quad \text{whenever } \theta_i = 0 \quad (1 \le i \le k)$$

$$\lim_{n\to\infty} P_{n,\theta}(\hat{\theta}_i = 0) = 0 \quad \text{whenever } \theta_i \ne 0 \quad (1 \le i \le k)$$

- Consistent vs. conservative model selection can in our context be driven by the asymptotic behavior of the tuning parameters λ_n. Also called "sparsely" vs. "non-sparsely" tuned procedures.

- Oracle property – Asymptotic distribution coincides with the one of the infeasible unpenalized estimator using the true zero restrictions (with VC-matrix $\Sigma_{\boldsymbol{\theta}}$).

$$n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \rightarrow N(\mathbf{0}, \Sigma_{\boldsymbol{\theta}})$$

  Seems to suggest that $\hat{\boldsymbol{\theta}}$ performs as well as if we would know the true zero coefficients of $\boldsymbol{\theta}$.

Literature on distributional properties of PLSEs

- Knight & Fu, 2000. Moving-parameter asymptotics for non-sparsely tuned LASSO and Bridge estimators in general.
- Fan & Li, 2001. Fixed-parameter asymptotics for SCAD.
- Zou, 2006. Fixed-parameter asymptotics for sparsely-tuned LASSO and adaptive LASSO.
- Additional papers establishing the oracle property for sparsely-tuned PLSEs and related estimators within a fixed-parameter framework.

  Fan & Li (2002, 2004), Bunea (2004), Bunea & McKeague (2005), Wang & Leng (2007), Li & Liang (2007), Wang, G. Li, & Tsai (2007), Zhang & Li (2007), Wang, R. Li, & Tsai (2007), Zou & Yuan (2008), Zou & Li (2008), Johnson, Lin, & Zeng (2008), ...

Literature on distributional properties of PLSEs (cont'd)

This talk is based on

- Pötscher & Leeb, 2007. Finite-sample distribution, moving-parameter asymptotics for hard-thresholding, LASSO, and SCAD. Impossibility result for the estimation of the cdf.
- Pötscher & Schneider, 2007. Analogous results for the adaptive LASSO.
- Pötscher & Schneider, 2008. Finite-sample and asymptotic coverage probabilities of confidence sets for hard-thresholing, LASSO, ad. LASSO.

Definition of the (adaptive) LASSO estimator $\hat{\theta}_{AL}$

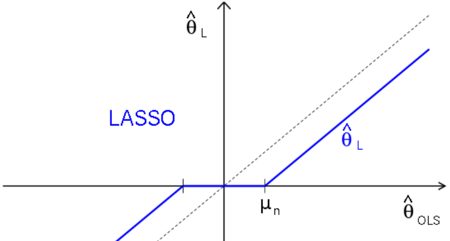
LASSO estimator (Tibshirani, 1996)

$$\hat{\boldsymbol{\theta}}_{L} = \underset{\boldsymbol{\theta}\in\mathbb{R}^k}{\arg\min} \ \|\mathbf{y} - X\boldsymbol{\theta}\|^2 + 2n\mu_n \sum_{i=1}^{k} |\theta_i| \qquad \mu_n > 0$$

Tuning parameter $\lambda_n = 2n\mu_n$. For $k = 1$: adaptive LASSO estimator (Zou, 2006)

$$\hat{\boldsymbol{\theta}}_{AL} = \underset{\boldsymbol{\theta}\in\mathbb{R}^k}{\arg\min} \ \|\mathbf{y} - X\boldsymbol{\theta}\|^2 + 2n\mu_n^2 \sum_{i=1}^{k} |\theta_i|/|\hat{\theta}_{OLS,j}| \quad \mu_n > 0$$

Tuning parameter $\lambda_n = 2n\mu_n^2$. For $k = 1$:



Two regimes for consistency

In terms of model selection consistency, two possible regimes for the tuning parameter μ_n arise.

1. The case $\mu_n \to 0$ and $n^{1/2}\mu_n \to m$, $0 \leq m < \infty$, corresponds to conservative model selection (non-sparsely tuned).
2. The case $\mu_n \to 0$ and $n^{1/2}\mu_n \to \infty$ corresponds to consistent model selection (sparsely tuned).

Remark (estimation consistency).
If $\mu_n \not\to 0$, then $\hat{\boldsymbol{\theta}}_{\text{AL}}$ is not even consistent for $\boldsymbol{\theta}$. Therefore, $\mu_n \to 0$ is a "basic condition".

We will focus on 2 here, also discuss 1 .

Asymptotics in the consistent case

Zou (2006) "oracle property"

Suppose $X'X/n \to Q > 0$ and $\varepsilon_t \overset{\text{iid}}{\sim} (0, \sigma^2)$.
If $\mu_n \to 0$ and $n^{1/2}\mu_n \to \infty$ and additionally $n^{1/4}\mu_n \to 0$, then

$$n^{1/2}(\hat{\boldsymbol{\theta}}_{\text{AL}} - \boldsymbol{\theta}) \to N(\mathbf{0}, \Sigma_\theta),$$

where Σ_θ is the asymptotic VC-matrix of the restricted LS-estimator based on the unknown true zero restrictions.

Questions

- Does this theorem provide meaningful insights? Finite-sample distribution?
- Asymptotic behavior under regime 1 ?
- What if condition $n^{1/4}\mu_n \to 0$ is dropped in 2 ?
- Pointwise vs. uniform consistency rates?
- Properties of confidence intervals?
- Estimability of finite-sample distribution?

We answer these questions within a normal linear regression model and address the non-orthogonal case in a simulation study.

Explicit solution in a simple model

- X is non-stochastic ($n \times k$), $rk(X) = k$.
- $\varepsilon \sim N_n(0, \sigma^2 \mathcal{I}_n)$
- For the theoretical analysis, assume that σ^2 is known and that $X'X$ is diagonal, in particular $X'X = n\mathcal{I}_k$.
- Remove these assumptions for simulation results concerning the finite-sample distribution.

Wlog consider Gaussian location model $y_1, \ldots, y_n \overset{\text{iid}}{\sim} N(\theta, 1)$.
Then $\hat{\theta}_{\text{OLS}} = \bar{y}$ with $\hat{\theta}_{\text{OLS}} \sim N(\theta, 1/n)$ and

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

Selects between restricted $\{N(0,1)\}$ and full model $\{N(\theta,1) : \theta \in \mathbb{R}\}$

The cdf $F_{n,\theta}(x) = P_{n,\theta}(n^{1/2}(\hat{\theta}_{\text{AL}} - \theta) \leq x)$ of $\hat{\theta}_{\text{AL}}$ is given by

$$\mathbf{1}(n^{1/2}\theta + x \geq 0)\, \Phi\left(z^{(2)}_{n,\theta}(x)\right) + \mathbf{1}(n^{1/2}\theta + x < 0)\, \Phi\left(z^{(1)}_{n,\theta}(x)\right).$$

$z^{(2)}_{n,\theta}(x)$ and $z^{(1)}_{n,\theta}(x)$ are $-(n^{1/2}\theta - x)/2 \pm \sqrt{((n^{1/2}\theta + x)/2)^2 + n\mu_n^2}$.

$dF_{n,\theta}$ is given by

$$\begin{aligned} &\left\{\, \Phi(n^{1/2}(-\theta + \mu_n)) \;-\; \Phi(n^{1/2}(-\theta - \mu_n)) \,\right\}\, d\delta_{-n^{1/2}\theta}(x) + \\ &0.5 \times \left\{ \mathbf{1}(n^{1/2}\theta + x > 0)\, \phi\left(z^{(2)}_{n,\theta}(x)\right)\, (1 + t_{n,\theta}(x)) + \right. \\ &\qquad \left. \mathbf{1}(n^{1/2}\theta + x < 0)\, \phi\left(z^{(1)}_{n,\theta}(x)\right)\, (1 - t_{n,\theta}(x)) \right\}\, dx \end{aligned}$$

where $t_{n,\theta}(x) := \left(((n^{1/2}\theta + x)/2)^2 + n\mu_n^2\right)^{-1/2}$.

Φ and ϕ the cdf and pdf of $N(0,1)$, resp.

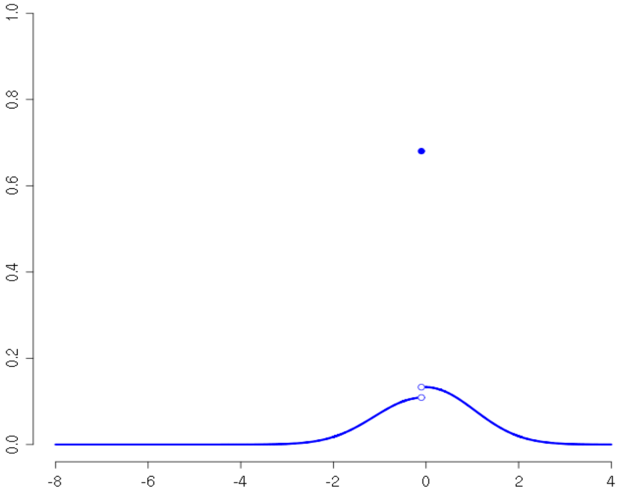
$n = 40$, $\theta = 0.05$, $\mu_n = 0.05$

Non-normality??

- Finite-sample distribution is highly non-normal.
- Oracle property predicts normality (asymptotically).

The Oracle (fixed-parameter asymptotics)

$n = 1, \qquad \mu_n = n^{-1/3}$ (consistent case)



$n = 10, \qquad \mu_n = n^{-1/3}$ (consistent case)

Moving-parameter asymptotics?

Let underlying parameter θ depend on sample size:

Let $\theta_n \in \mathbb{R}$ be arbitrary, subject only to
$\theta_n/\mu_n \to \zeta \in \mathbb{R} \cup \{-\infty, \infty\}$ and $n^{1/2}\theta_n \to \nu \in \mathbb{R} \cup \{-\infty, \infty\}$.

This is not really a restriction since every subsequence of θ_n contains a further subsequence with these properties. Also note that $\zeta \neq 0$ implies $\nu = \pm\infty$.

Moving-parameter asymptotics ②

② Consistent case.

Let $\mu_n \to 0$ and $n^{1/2}\mu_n \to \infty$. Suppose the true parameter $\theta_n \in \mathbb{R}$ satisfies $\theta_n/\mu_n \to \zeta \in \mathbb{R} \cup \{-\infty, \infty\}$ and $n^{1/2}\theta_n \to \nu \in \mathbb{R} \cup \{-\infty, \infty\}$. Then F_{n,θ_n} converges weakly to

- If $0 \leq |\zeta| < \infty$: pointmass at $-\nu$
- If $|\zeta| = \infty$: $\Phi(. + \rho/\theta)$ where $n^{1/2}\mu_n^2 \to \rho$.

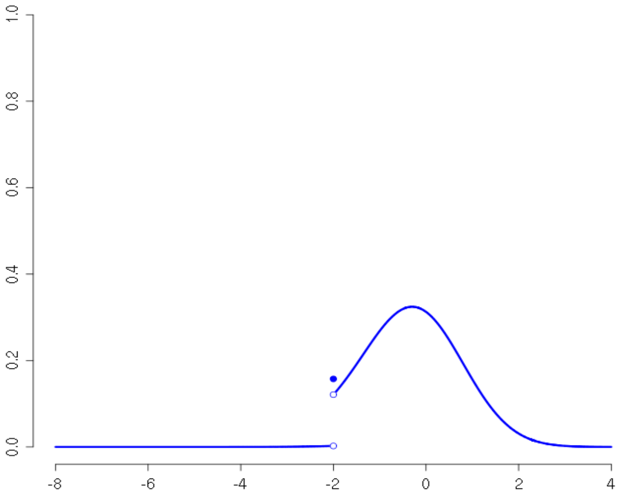
Depending on ζ, ν and ρ, three possible limits arise.

- Distribution collapses at a point.
- Total mass escapes to $\pm\infty$.
- Limit distribution is normal (possibly shifted!).

Non-normality persists!!

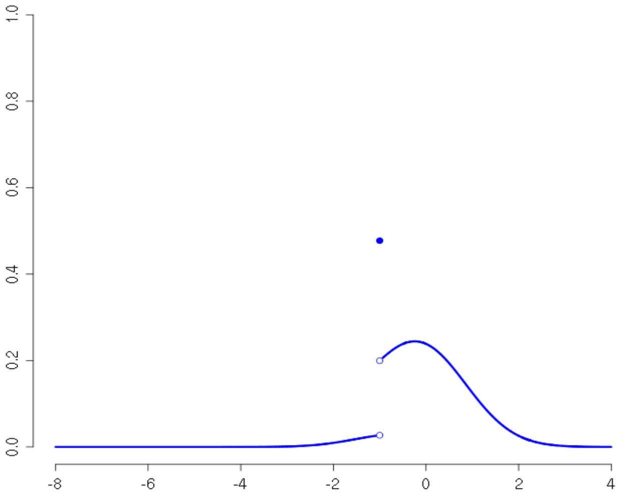
Illustration: collapsing to pointmass

Example 1: $n = 1$, $\zeta = 0,\ \nu = 2$ $(\mu_n = n^{-1/3},\ \theta_n = 2n^{-1/2})$



Example 1: $n = 10$, $\zeta = 0,\ \nu = 2$ $(\mu_n = n^{-1/3},\ \theta_n = 2n^{-1/2})$

Example 2: $n = 1, \quad \zeta = 1, \ \nu = \infty \quad (\mu_n = n^{-1/5}, \ \theta_n = n^{-1/5})$



Example 2: $n = 10, \quad \zeta = 1, \ \nu = \infty \quad (\mu_n = n^{-1/3}, \ \theta_n = n^{-1/5})$

② Consistent case.

Let $\mu_n \to 0$ and $n^{1/2}\mu_n \to \infty$. Suppose the true parameter $\theta_n \in \mathbb{R}$ satisfies $\theta_n/\mu_n \to \zeta \in \mathbb{R} \cup \{-\infty, \infty\}$ and $n^{1/2}\theta_n \to \nu \in \mathbb{R} \cup \{-\infty, \infty\}$. Then F_{n,θ_n} converges weakly to

- If $0 \leq |\zeta| < \infty$: pointmass at $-\nu$
- If $|\zeta| = \infty$: $\Phi(. + \rho/\theta)$ where $n^{1/2}\mu_n^2 \to \rho$.

Zou (pointwise case) ? Above theorem implies that

$$F_{n,\theta}(x) \to \begin{cases} \mathbf{1}(x \geq 0) & \theta = 0 \quad (\implies \zeta, \nu = 0) \\ \Phi(x + \rho/\theta) & \theta \neq 0 \quad (\implies |\zeta| = \infty) \end{cases}$$

Remark: $\rho = 0 \iff n^{1/4}\mu_n \to 0$.

Uniform consistency and alternative scaling ②

- Adaptive LASSO has in a uniform sense a rate of convergence that is slower than $n^{1/2}$.
- The "correct" uniform rate can be shown to be μ_n^{-1}.
- In a moving-parameter framework, the asymptotic distribution of $\mu_n^{-1}(\hat{\theta}_{\mathrm{AL}} - \theta)$ collapses to pointmass.

Let $\mu_n \to 0$ and $n^{1/2}\mu_n \to \infty$. Suppose the true parameter $\theta_n \in \mathbb{R}$ satisfies $\theta_n/\mu_n \to \zeta \in \mathbb{R} \cup \{-\infty, \infty\}$. Then $G_{n,\theta_n} := P(\mu_n^{-1}(\hat{\theta}_{\mathrm{AL}} - \theta) \leq x)$ converges weakly to

- If $|\zeta| < 1$: pointmass at $-\zeta$
- If $1 \leq |\zeta| < \infty$: pointmass at $-1/\zeta$
- If $|\zeta| = \infty$: pointmass at 0

Above theorems reflect that

$$\hat{\theta}_{\mathsf{AL}} - \theta = \text{“}\textsc{bias}\text{”} + \text{“}\textsc{fluctuation}\text{”}$$

where

- "BIAS" is $O(n^{-1/2})$ in a pointwise sense but is only $O(\mu_n)$ in a uniform sense, whereas
- "FLUCTUATION" is always of order $n^{-1/2}$.

Moving parameter asymptotics

1. Conservative case.

Let $\mu_n \to 0$ and $n^{1/2}\mu_n \to m,\ 0 \le m < \infty$. Suppose the true parameter $\theta_n \in \mathbb{R}$ satisfies $n^{1/2}\theta_n \to \nu \in \mathbb{R} \cup \{-\infty, \infty\}$. Then F_{n,θ_n} converges weakly to

- If $\nu \in \mathbb{R}$
$$\mathbf{1}(\nu + x \ge 0)\, \Phi\left(-(\nu - x)/2 + \sqrt{((\nu + x)/2)^2 + m^2}\right) + \mathbf{1}(\nu + x < 0)\, \Phi\left(-(\nu - x)/2 - \sqrt{((\nu + x)/2)^2 + m^2}\right)$$
- $\Phi(x)$ if $|\nu| = \infty$.

Note: Asymptotic distributions are the same as finite-sample distribution, except that $n^{1/2}\theta_n$ and $n^{1/2}\mu_n$ have settled down to their limiting values, capturing finite-sample behavior very well.

Moving parameter asymptotics (1) (cont'd)

- $\hat{\theta}_{\text{AL}}$ is now uniformly $n^{1/2}$-consistent.
- Fixed-parameter asymptotics: previous theorem implies that $F_{n,\theta}(x)$ converges to
  - $\mathbf{1}(x \geq 0)\,\Phi\left(\frac{x}{2} + \sqrt{(\frac{x}{2})^2 + m^2}\right) + \mathbf{1}(x < 0)\,\Phi\left(\frac{x}{2} - \sqrt{(\frac{x}{2})^2 + m^2}\right)$

    if $\theta = 0$ $(\nu = 0)$
  - $\Phi(x)$ if $\theta \neq 0$ $(|\nu| = \infty)$
- Fixed-parameter asymptotic distributions are also non-normal, capturing behavior the finite-sample distributions to some extent (no oracle here).

Other PLSEs

Results are similar for hard-thresholding, soft-thresholding (LASSO), and SCAD estimator. (Pötscher & Leeb, 2007).

- Identical results in terms of (uniform) consistency.
- Analogous (asymptotic) distributional results.

Confidence sets based on the adaptive LASSO

Let $C_n = [\hat{\theta}_{\text{AL}} - a_n, \hat{\theta}_{\text{AL}} + b_n]$.

The infimal coverage probability $\inf_{\theta \in \mathbb{R}} P_{n,\theta}(\hat{\theta}_{\text{AL}} \in C_n)$ is given by

$$\Phi(n^{1/2}(a_n - \mu_n)) - \Phi\left(n^{1/2}((a_n - b_n)/2 - \sqrt{((a_n + b_n)/2)^2 + \mu_n^2}\right)$$

if $a_n \leq b_n$ and

$$\Phi\left(n^{1/2}((a_n - b_n)/2 + \sqrt{((a_n + b_n)/2)^2 + \mu_n^2})\right) - \Phi(n^{1/2}(-b_n + \mu_n))$$

if $a_n > b_n$.

Symmetric intervals ($a_n = b_n$) can be shown to be the shortest ones for a given infimal coverage probability δ.

Confidence sets based on PLSEs

- For each $n \in \mathbb{N}$, we have

$$a_{n,H} > a_{n,\mathrm{AL}} > a_{n,L} > a_{n,\mathrm{OLS}} \quad \text{for a given } \delta > 0$$

- Asymptotically, the following holds.

1. Conservative case. All quantities are of the same order $n^{-1/2}$.

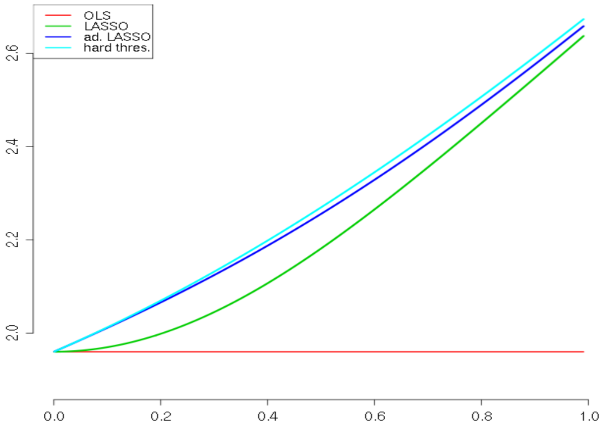
$$a_{n,H} \sim a_{n,\mathrm{AL}} \sim a_{n,L} \sim a_{\mathrm{OLS}}$$

2. Consistent case. $a_{n,H}, a_{n,L}$, and $a_{n,A}$ are one order of magnitude larger than $a_{n,\mathrm{OLS}}$.

$$a_H / a_{\mathrm{OLS}} \sim a_{\mathrm{AL}} / a_{\mathrm{OLS}} \sim a_L / a_{\mathrm{OLS}} \sim n^{1/2} \mu_n \to \infty$$

Confidence sets based on PLSEs (cont'd)

Plot of $n^{1/2}a_n$ against $n^{1/2}\mu_n$ for $\delta = 0.95$.

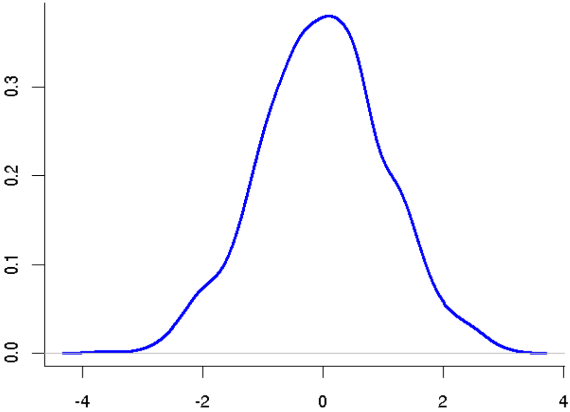


Simulations - remove orthogonality assumption

$k = 4$, $n = 200$, $\theta = (3, 1.5, 0, 0)' + 2/n^{1/2}(0, 0, 1, 1)'$, $X'X = n\Omega$ with $\Omega_{ij} = 0.5^{|i-j|}$, 1000 simulations

- $\mu_n = n^{-1/3}$

θ_1



θ_2

Estimation of the cdf of $n^{1/2}(\hat{\theta}_{\text{AL}} - \theta)$?

Let $F_{n,\theta}$ be the distribution function of $n^{1/2}(\hat{\theta}_{\text{AL}} - \theta)$.

Let $\mu_n \to 0$ and $n^{1/2}\mu_n \to m$ with $0 \leq m \leq \infty$. Then *every consistent* estimator $\hat{F}_n(t)$ of $F_{n,\theta}(t)$ satisfies

$$\lim_{n\to\infty} \sup_{|\theta|<c/n^{1/2}} P_{n,\theta}\left(\left|\hat{F}_n(t) - F_{n,\theta}(t)\right| > \varepsilon\right) \geq \frac{1}{2}$$

for each $\varepsilon < (\Phi(t+m) - \Phi(t-m))/2$ and each $c > |t|$.

In particular, not uniformly consistent estimator for $F_{n,\theta}(t)$ exisits!

Analogous result for cdf under μ^{-1}-scaling.

Proof rests on Pötscher & Leeb (2006).

Finite-sample result:

Let $\mu_n \to 0$ and $n^{1/2}\mu_n \to m$ with $0 \leq m \leq \infty$. Then *every* estimator $\hat{F}_n(t)$ of $F_{n,\theta}(t)$ satisfies

Conclusions

- The finite-sample distribution of the adaptive LASSO estimator and other PLSEs are highly non-normal.
- Non-normality persists in large samples. This can be seen through a "moving-parameter" asymptotic framework.
- Fixed-parameter asymptotics (as underlying the oracle-property) paint a misleading picture of the performance of the estimator due to the non-uniformity of these results.
- Confidence intervals in the consistent case are larger by one order of magnitude compared to the unpenalized estimator.
- The distribution function of the adaptive LASSO estimator and other PLSEs cannot be estimated in a uniformly consistent manner.
- NOT a critisim on PLSEs per se, but relying on fixed-parameter asymptotics in this context is dangerous.

References

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Ass.*, 96:1348–1360, 2001.

I. E. Frank and J. H. Friedman. A statistical view of some chemometrics regression tools (with discussion). *Technom.*, 35:109–148, 1993.

A.E. Hoerl and R. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12, 55–67.

K. Knight and W. Fu. Asymptotics of lasso-type estimators. *Ann. Stat.*, 28:1356–1378, 2000.

H. Leeb and B. M. Pötscher. Performance limits for estimators of the risk or distribution of shrinkage-type estimators, and some general lower risk-bound results. *Economet. Theory*, 22:69–97, 2006. Corrections. Ibidem, 24:581–583, 2008.

B. M. Pötscher. Confidence sets based on sparse estimators are necessarily large. *Manuscript*, 2007. arXiv:0711.1036.

B. M. Pötscher and H. Leeb. On the distribution of penalized maximum likelihood estimators: The LASSO, SCAD, and thresholding. *Manuscript*, 2007. arXiv:0711.0660.

B. M. Pötscher and U. Schneider. Confidence sets based on penalized maximum likelihood estimators. *Manuscript*, 2008. arXiv:0806.1652.

B. M. Pötscher and U. Schneider. On the distribution of the adaptive lasso estimator. *J. Stat. Plan. Inf.*, to appear.

R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Stat. Soc. B*, 58:267–288, 1996.

H. Zou. The adaptive lasso and its oracle properties. *J. Am. Stat. Ass.*, 101:1418–1429, 2006.