On the Distribution of the Adaptive LASSO Estimator

U. Schneider
(joint with B. M. Pötscher)

Universität Wien

Workshop on Current Trends and Challenges in Model Selection,
Vienna, July 24, 2008

Penalized ML Estimators

Linear regression model $y = X\theta + u$, consider estimator $\hat{\theta}$ for θ

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

λ_n is a tuning parameter.

- Bridge estimators (l_p - type penalties, Frank and Friedman, 1993, LASSO for $p = 1$, Tibshirani, 1996).
- Hard- and soft-thresholding estimators.
- Smoothly clipped absolute deviation (SCAD) estimator (Fan and Li, 2001).
- Adaptive LASSO estimator (Zou, 2006).

These estimators can be viewed to simultaneously perform model selection and parameter estimation. ($p \leq 1$ for Bridge est.)

Some terminology

- Conservative model selection – Zero coefficients are found with asymptotic probability less than 1.
- Consistent model selection – Zero coefficients are found with asymptotic probability equal to 1.
- Oracle property – Asymptotic distribution coincides with the one of the unpenalized estimator of the true model.

Consistent vs. conservative model selection is in our context driven by the asymptotic choice of tuning parameters λ_n. ("Sparsely" vs. "non-sparsely" tuned procedures).

Some literature on distributional properties of PMLEs

- Knight and Fu, 2000. Moving-parameter asymptotics for (non-sparsely tuned) LASSO and Bridge estimators in general.
- Fan and Li, 2001. Fixed-parameter asymptotics for SCAD.
- Zou, 2006. Fixed-parameter asymptotics for LASSO and adaptive LASSO.
- Pötscher and Leeb, 2007. Finite-sample distribution, moving-parameter asymptotics for hard-thresholding, LASSO, and SCAD. Impossibility result for the estimation of the cdf.
- Pötscher and Schneider, 2007. Analogous results for the adaptive LASSO.
- Pötscher and Schneider, 2008. Finite-sample and asymptotic coverage probabilities of confidence sets for hard-thresholing, LASSO, ad. LASSO.
- ...

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

Linear regression model $y = X\theta + u$.

- X is $n \times k$, non-stochastic, $\text{rk}(X) = k$.
- $u \sim N_n\left(0, \sigma^2 \mathcal{I}_n\right)$

Adaptive LASSO estimator, Zou, 2006 (random penalty weights)

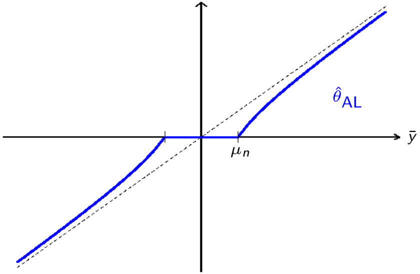
$$\hat{\theta}_{\text{AL}} = \underset{\theta \in \mathbb{R}^k}{\arg\min}\ \|y - X\theta\|^2 + 2n\mu_n^2 \sum_{j=1}^{k} |\theta_j| / |\hat{\theta}_{\text{OLS},j}|, \quad \mu_n > 0$$

- 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.

Explicit solution in the simplified model

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

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



Consistency of $\hat{\theta}_{\text{AL}}$

- Estimation consistency:
  - The condition $\mu_n \to 0$ is equivalent to the consistency of $\hat{\theta}_{\text{AL}}$.
  - Then $\hat{\theta}_{\text{AL}}$ is also is also uniformly consistent for θ, i.e. for all $\varepsilon > 0$

$$\lim_{n\to\infty} \sup_{\theta\in\mathbb{R}} P_{n,\theta}\left(\left|\hat{\theta}_{\text{AL}} - \theta\right| > \varepsilon\right) = 0$$

- Model selection consistency: two possible regimes arise.
  1. The case $\mu_n \to 0$ and $n^{1/2}\mu_n \to m$, $0 \le m < \infty$, 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).

The finite-sample distribution of $\hat{\theta}_{\mathrm{AL}}$

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

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

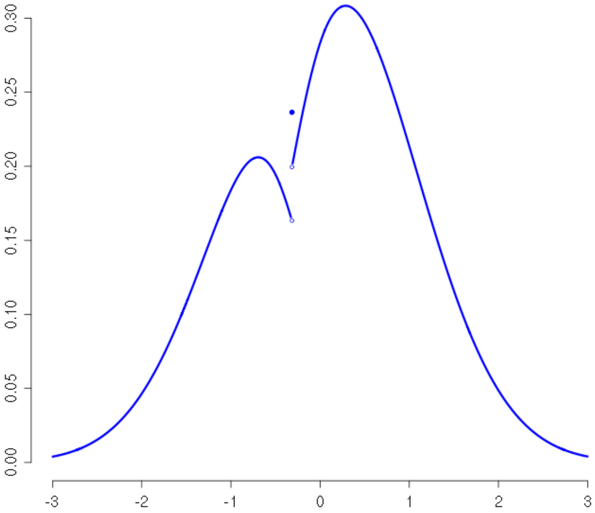
$z^{(2)}_{n,\theta}(x)$ and $z^{(1)}_{n,\theta}(x)$ are $-(n^{1/2}\theta - x)/2 \pm \sqrt{((n^{1/2}\theta + x)/2)^2 + n\mu_n^2}$.

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

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.

The finite-sample distribution of $\hat{\theta}_{AL}$

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



Fixed-parameter asymptotics – both regimes

1. Conservative case. $F_{n,\theta}$ converges weakly to

$$\begin{cases} \mathbf{1}(x \geq 0)\,\Phi\left(\frac{x}{2} + \sqrt{(\frac{x}{2})^2 + m^2}\right) + \mathbf{1}(x < 0)\,\Phi\left(\frac{x}{2} - \sqrt{(\frac{x}{2})^2 + m^2}\right) & \theta = 0 \\ \Phi(x) & \theta \neq 0 \end{cases}$$

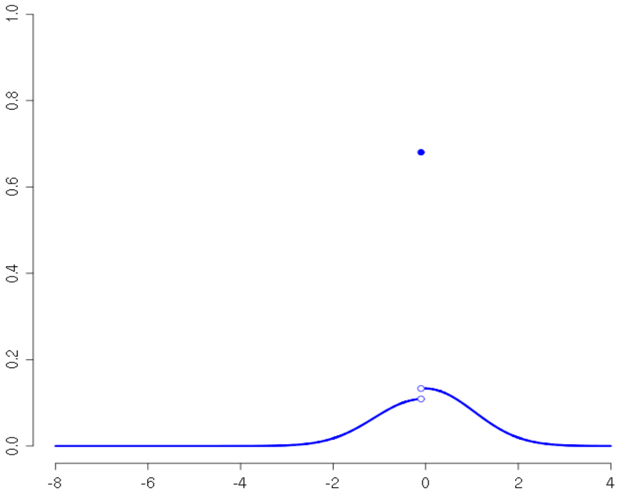
2. Consistent case. $F_{n,\theta}$ converges weakly to

$$\begin{cases} \mathbf{1}(x \geq 0) & \theta = 0 \\ \Phi(x + \rho\theta) & \theta \neq 0 \text{ and } n^{1/2}\mu_n^2 \to \rho \end{cases}$$

If $n^{1/4}\mu_n \to 0$, $F_{n,\theta}(x) \to \Phi(x)$ for $\theta \neq 0$ ("oracle property", Zou, 2006).

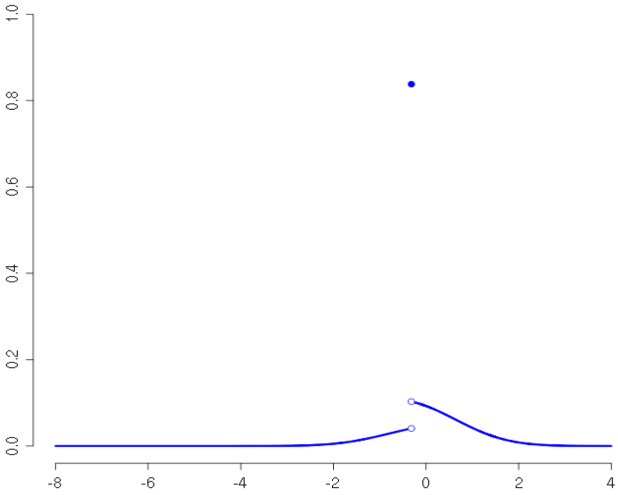
Fixed-parameter asymptotic – consistent case

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



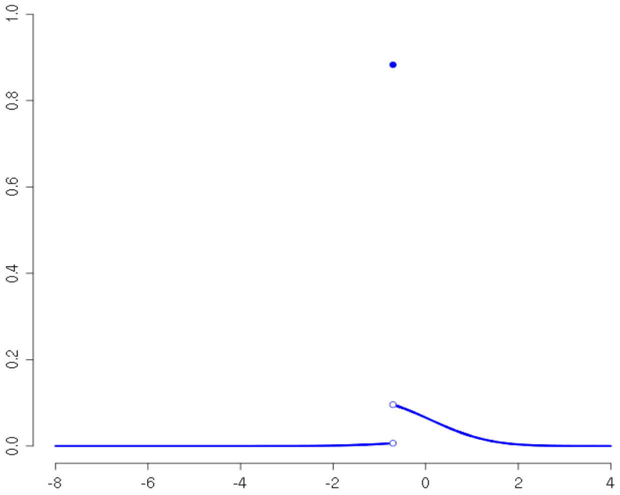
Fixed-parameter asymptotic – consistent case

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



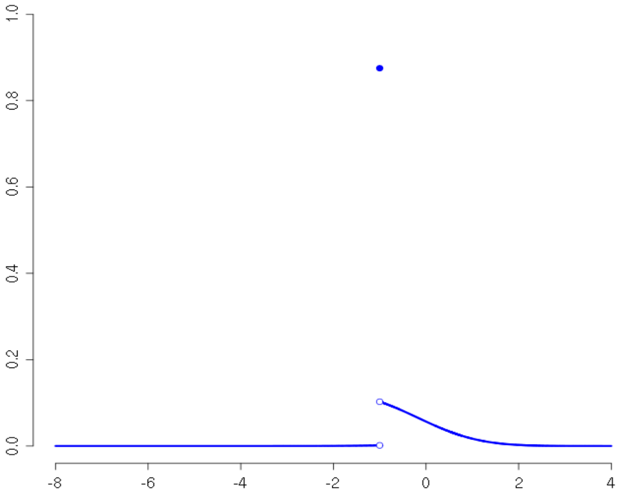
Fixed-parameter asymptotic – consistent case

$n = 50, \qquad \mu_n = n^{-1/3}$ (consistent case)



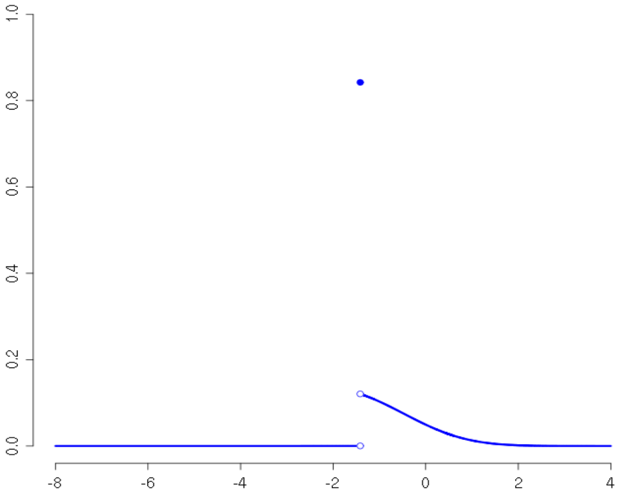
Fixed-parameter asymptotic – consistent case

$n = 100, \qquad \mu_n = n^{-1/3}$ (consistent case)



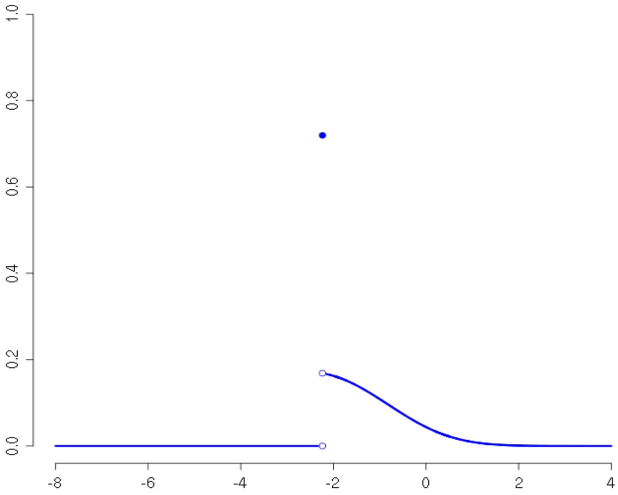
Fixed-parameter asymptotic – consistent case

$$n = 200, \qquad \mu_n = n^{-1/3} \text{ (consistent case)}$$



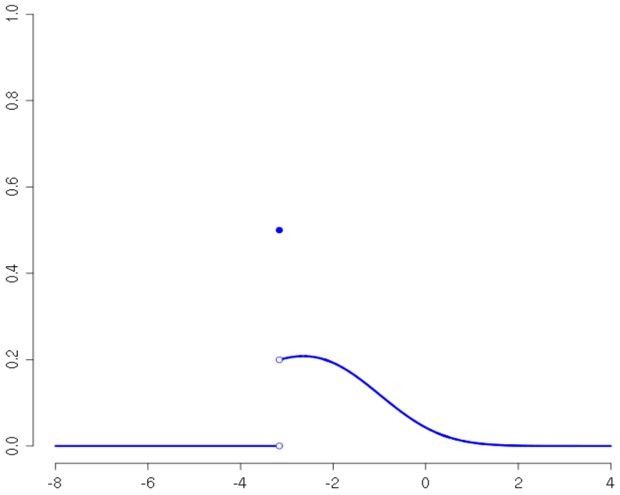
Fixed-parameter asymptotic – consistent case

$$n = 500, \qquad \mu_n = n^{-1/3} \text{ (consistent case)}$$



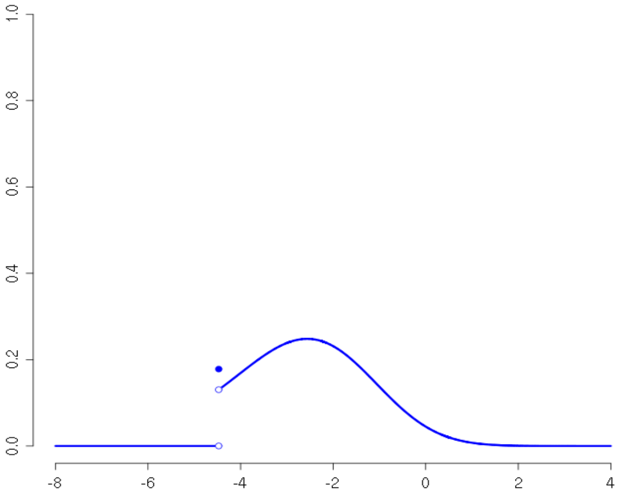
Fixed-parameter asymptotic – consistent case

$$n = 1000, \qquad \mu_n = n^{-1/3} \text{ (consistent case)}$$



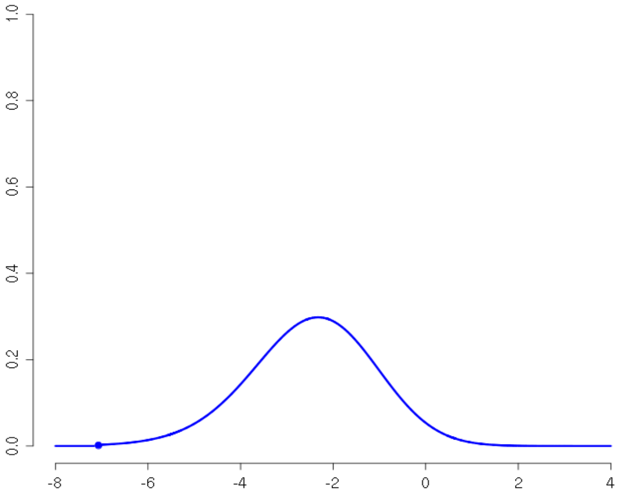
Fixed-parameter asymptotic – consistent case

$$n = 2000, \qquad \mu_n = n^{-1/3} \text{ (consistent case)}$$



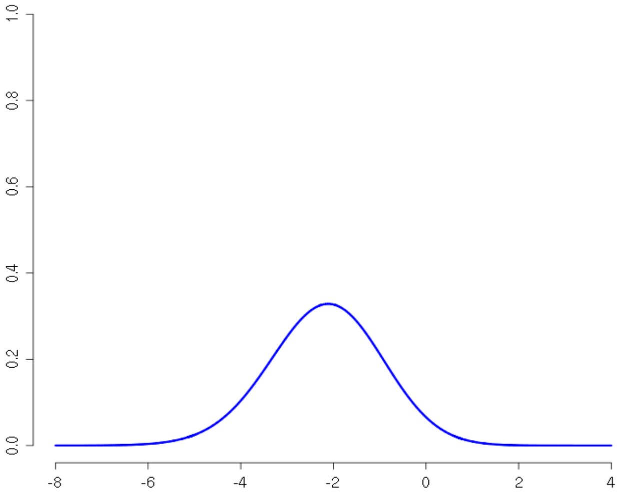
Fixed-parameter asymptotic – consistent case

$n = 5000, \qquad \mu_n = n^{-1/3}$ (consistent case)



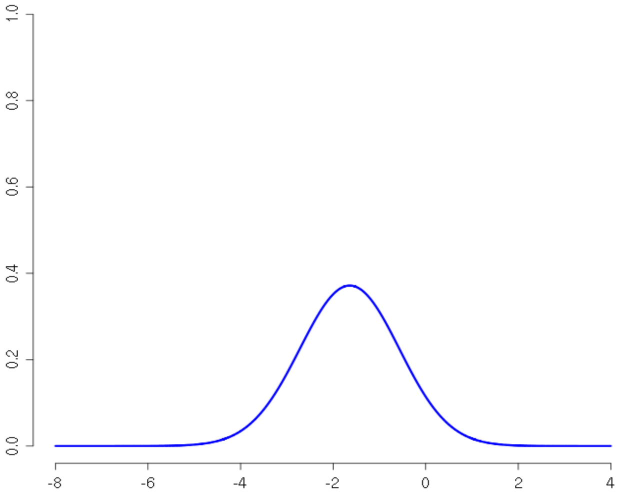
Fixed-parameter asymptotic – consistent case

$n = 10^4, \qquad \mu_n = n^{-1/3}$ (consistent case)



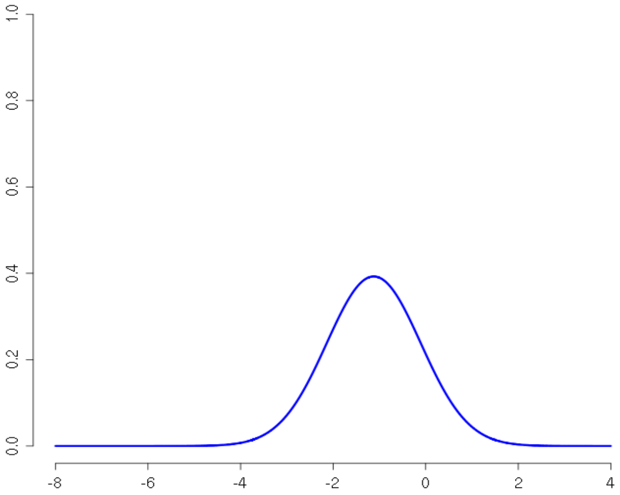
Fixed-parameter asymptotic – consistent case

$$n = 5 \times 10^4, \; \mu_n = n^{-1/3} \text{ (consistent case)}$$



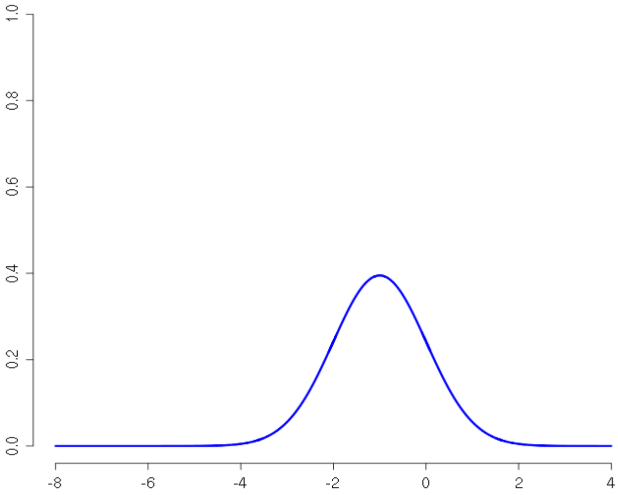
Fixed-parameter asymptotic – consistent case

$$n = 5 \times 10^5,\ \mu_n = n^{-1/3}\ \text{(consistent case)}$$



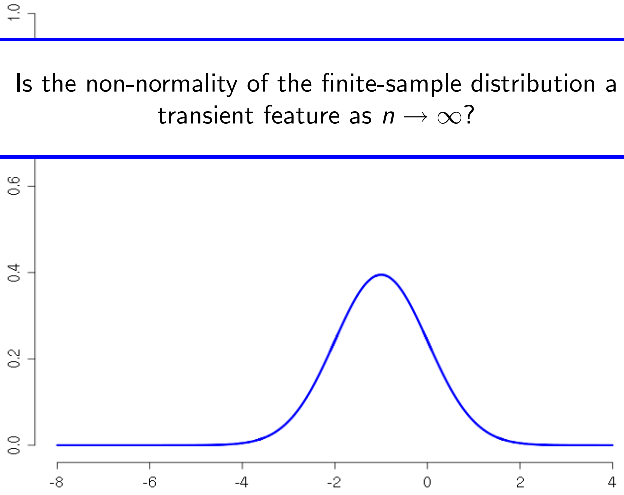
Fixed-parameter asymptotic – consistent case

$$n = 10^6, \qquad \mu_n = n^{-1/3} \text{ (consistent case)}$$



Fixed-parameter asymptotic – consistent case

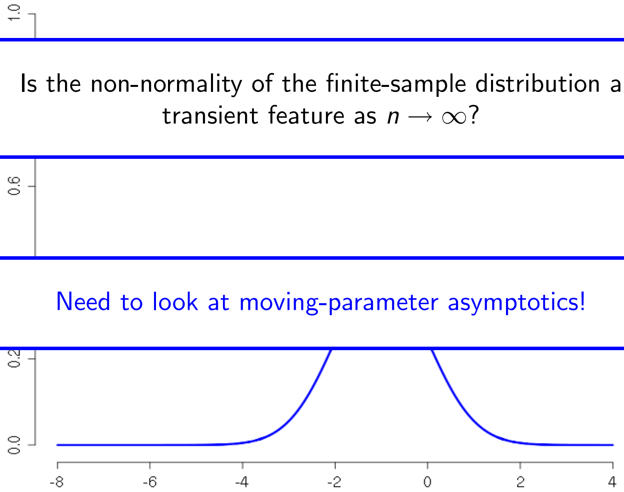
$$n = 10^6, \qquad \mu_n = n^{-1/3} \text{ (consistent case)}$$



Is the non-normality of the finite-sample distribution a transient feature as $n \to \infty$?

Fixed-parameter asymptotic – consistent case

$$n = 10^6, \qquad \mu_n = n^{-1/3} \text{ (consistent case)}$$



Is the non-normality of the finite-sample distribution a transient feature as $n \to \infty$?

Need to look at moving-parameter asymptotics!

Moving-parameter asymptotics

1. Conservative case.

Let $\mu_n \to 0$ and $n^{1/2}\mu_n \to m$, $0 \leq m < \infty$. 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_{A,n,θ_n} converges weakly to

- *If $\nu \in \mathbb{R}$*

$$\mathbf{1}(\nu + x \geq 0)\, \Phi\left(-(\nu - x)/2 + \sqrt{((\nu + x)/2)^2 + m^2}\right) +$$
$$\mathbf{1}(\nu + x < 0)\, \Phi\left(-(\nu - x)/2 - \sqrt{((\nu + x)/2)^2 + m^2}\right)$$

- $\Phi(x)$ *if $|\nu| = \infty$.*

Note: 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.

Moving-parameter asymptotics

1. 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_{A,n,θ_n} converges weakly to

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

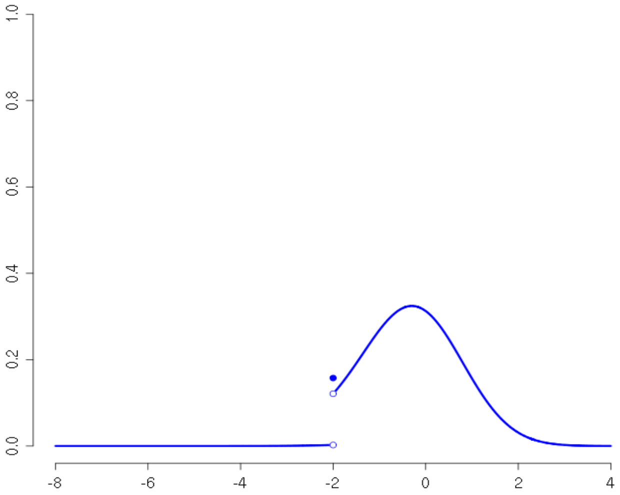
For $|\nu|, |\rho| = \infty$, above expressions mean total mass escaping to $\pm\infty$. Depending on ζ and ν, three possible (weak) limits arise.

- Distribution collapses at a point.
- Total mass escapes to $\pm\infty$.
- Limit distribution is normal.

Non-normality persists!!

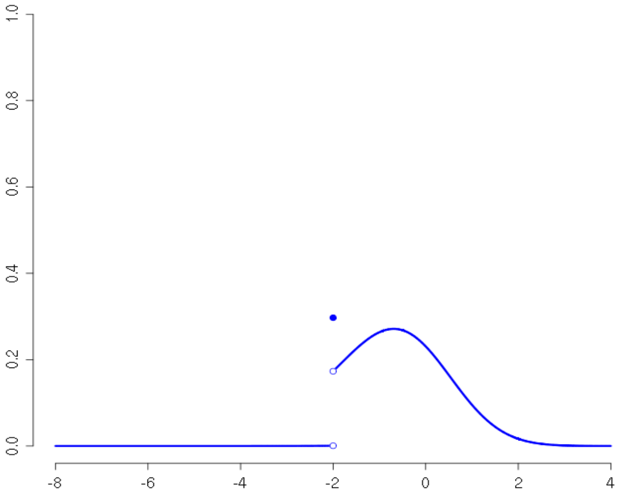
Moving-parameter asymptotics – consistent case

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



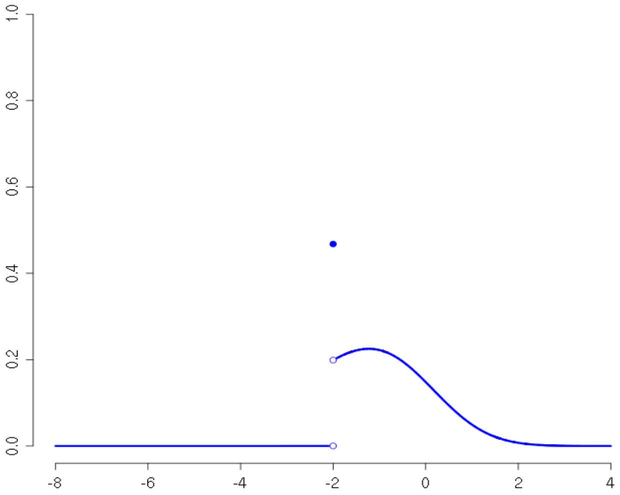
Moving-parameter asymptotics – consistent case

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



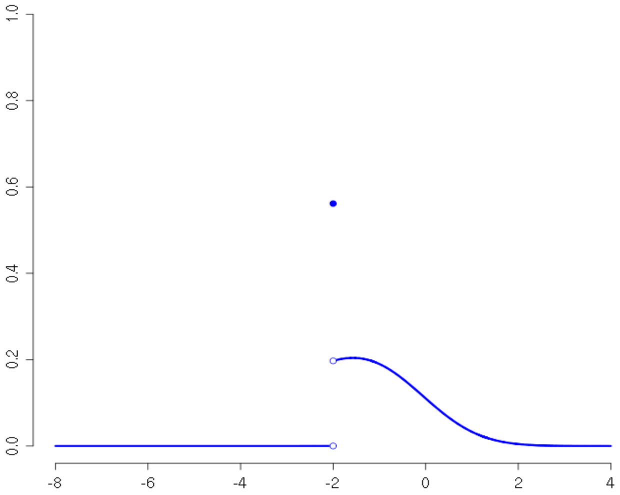
Moving-parameter asymptotics – consistent case

$$n = 50, \qquad \zeta = 0,\ \nu = 2 \qquad (\mu_n = n^{-1/3},\ \theta_n = 2n^{-1/2})$$



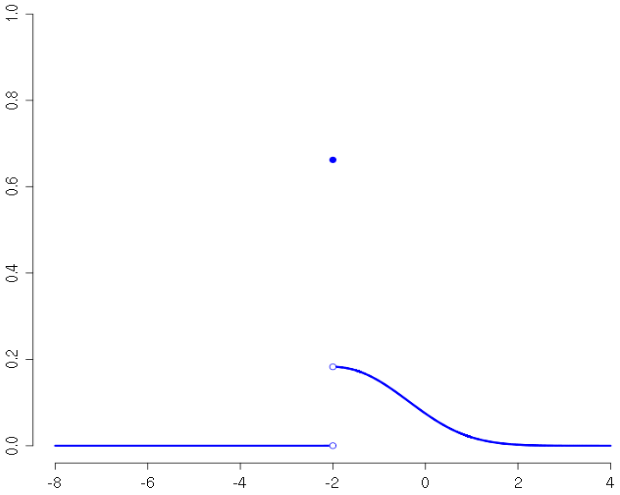
Moving-parameter asymptotics – consistent case

$$n = 100, \qquad \zeta = 0,\ \nu = 2 \qquad (\mu_n = n^{-1/3},\ \theta_n = 2n^{-1/2})$$



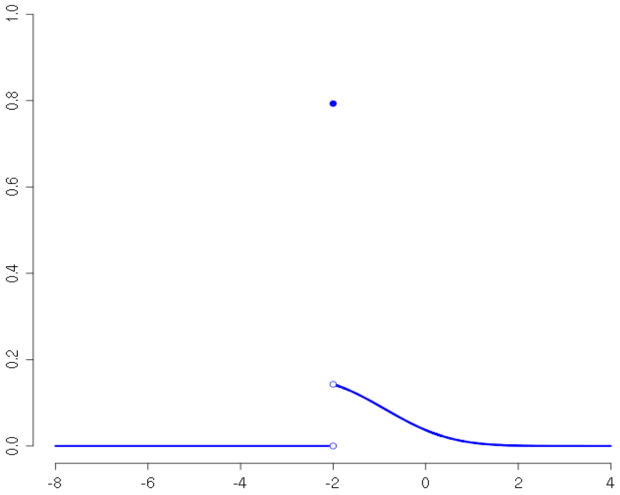
Moving-parameter asymptotics – consistent case

$$n = 200, \qquad \zeta = 0,\ \nu = 2 \qquad (\mu_n = n^{-1/3},\ \theta_n = 2n^{-1/2})$$



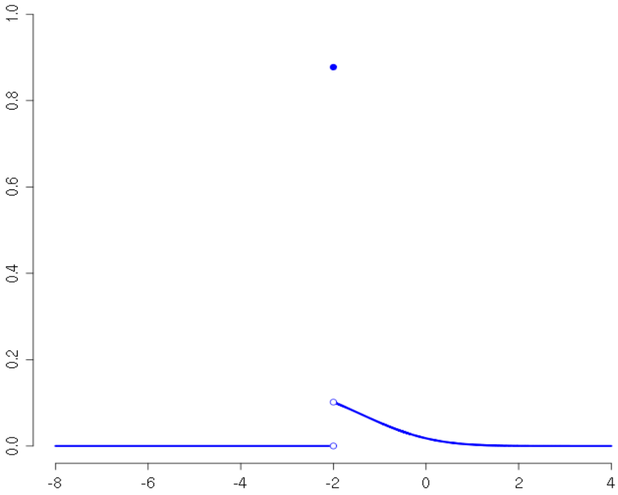
Moving-parameter asymptotics – consistent case

$$n = 500, \qquad \zeta = 0,\ \nu = 2 \qquad (\mu_n = n^{-1/3},\ \theta_n = 2n^{-1/2})$$



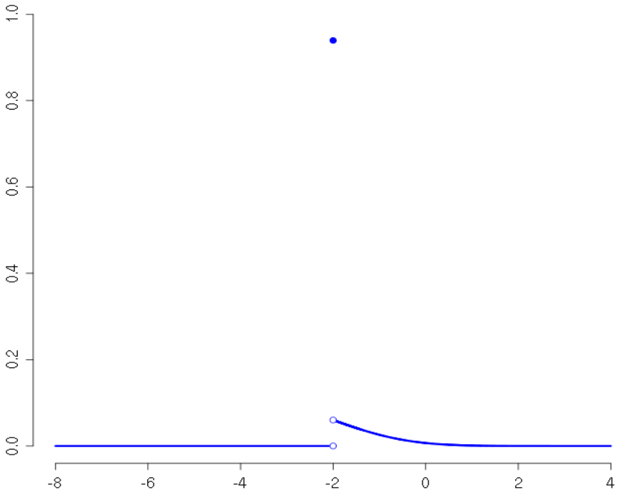
Moving-parameter asymptotics – consistent case

$$n = 1000, \qquad \zeta = 0,\ \nu = 2 \qquad (\mu_n = n^{-1/3},\ \theta_n = 2n^{-1/2})$$



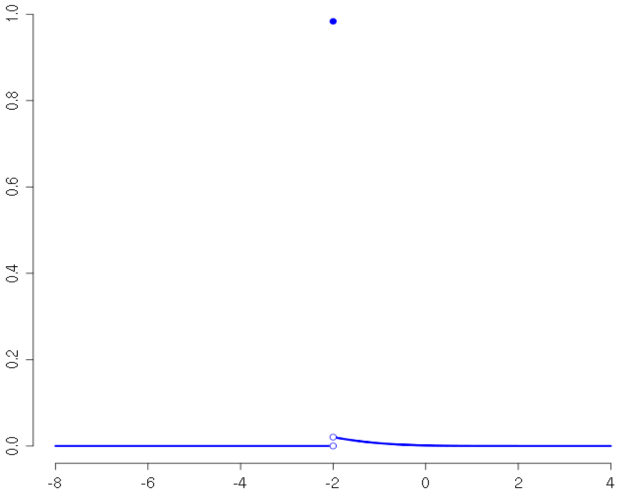
Moving-parameter asymptotics – consistent case

$$n = 2000, \qquad \zeta = 0,\ \nu = 2 \qquad (\mu_n = n^{-1/3},\ \theta_n = 2n^{-1/2})$$



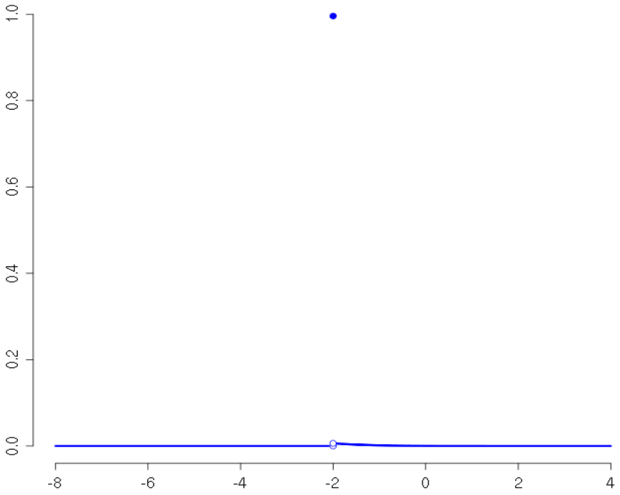
Moving-parameter asymptotics – consistent case

$$n = 5000, \qquad \zeta = 0,\ \nu = 2 \qquad (\mu_n = n^{-1/3},\ \theta_n = 2n^{-1/2})$$



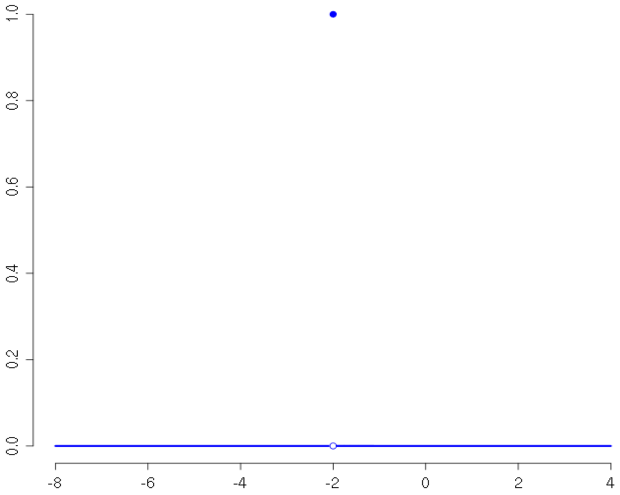
Moving-parameter asymptotics – consistent case

$$n = 10^4, \qquad \zeta = 0,\ \nu = 2 \qquad (\mu_n = n^{-1/3},\ \theta_n = 2n^{-1/2})$$



Moving-parameter asymptotics – consistent case

$$n = 5 \times 10^4, \ \ \zeta = 0, \ \nu = 2 \qquad (\mu_n = n^{-1/3}, \ \theta_n = 2n^{-1/2})$$



Uniform consistency with rate a_n

For which rate a_n is $n^{1/2}(\hat{\theta}_{\text{AL}} - \theta)$ uniformly a_n-consistent, i.e.

$$\lim_{M \to \infty} \sup_{n \in \mathbb{N}} \sup_{\theta \in \mathbb{R}} P_{n,\theta}\left(a_n \left|\hat{\theta}_{\text{AL}} - \theta\right| > M\right) = 0 \ ??$$

1. Conservative case. Rate a_n is $O(n^{1/2})$ (see prev. theorem).
2. Consistent case. Rate a_n is only $O(\mu_n^{-1})$.

   (In a moving-parameter framework, the asymptotic distribution of $\mu_n^{-1}(\hat{\theta}_{\text{AL}} - \theta)$ collapses to pointmass.)

Other PMLEs

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

- Identical consistency results.
- Analogous asymptotic results.

Confidence sets based on PMLEs

Based on Pötscher and Schneider, 2008.

Let $C_n = [\hat{\theta} - a_n, \hat{\theta} + a_n]$ be a confidence set for θ with infimal coverage probability of at least δ, ie $\inf_{\theta \in \mathbb{R}} P_{n,\theta}(\theta \in C_n) \geq \delta$.

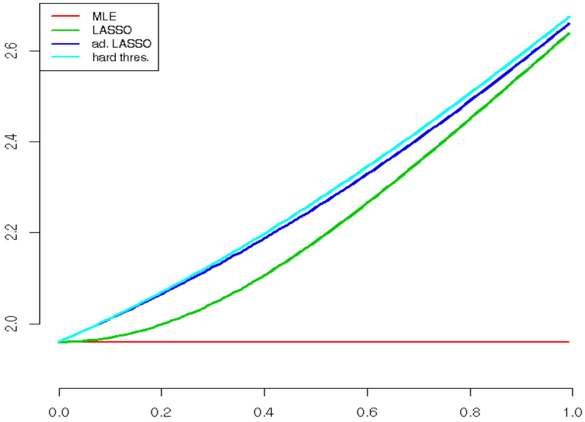
- For each $n \in \mathbb{N}$, we have

$$a_{n,H} > a_{n,L} > a_{n,A} > a_{n,\mathrm{MLE}} \text{ for a given } \delta > 0$$

- Asymptotically, the following holds.
  1. Conservative case. All quantities are of the same order $n^{-1/2}$.
  2. Consistent case. $a_{n,H}, a_{n,L}$, and $a_{n,A}$ are one order of magnitude larger than $a_{n,\mathrm{MLE}}$.

Confidence sets based on PMLEs

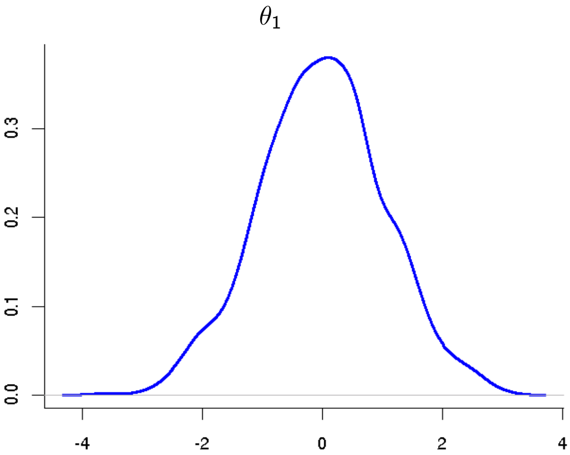
Plot of $n^{1/2}a_n$ against $n^{1/2}\mu_n$ for $\delta = 0.95$.



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

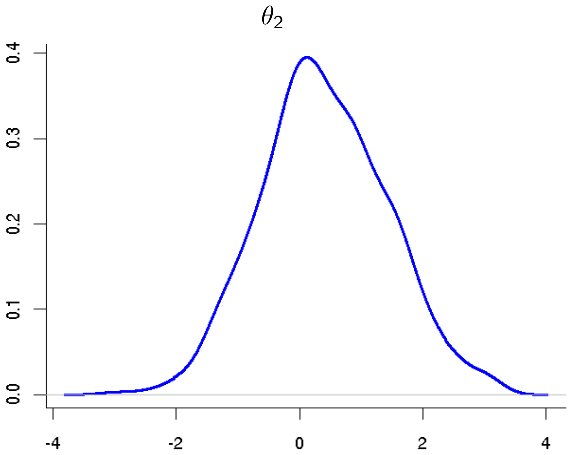
- $\mu_n = n^{-1/3}$



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

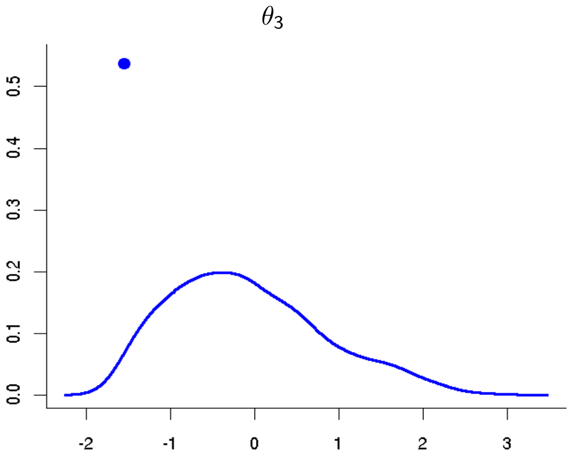
- $\mu_n = n^{-1/3}$



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

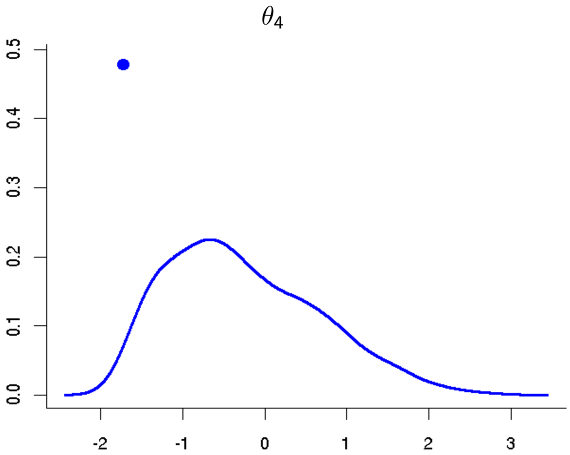
- $\mu_n = n^{-1/3}$



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

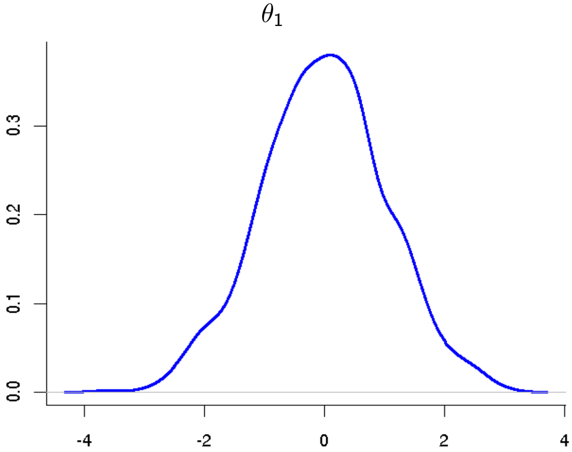
- $\mu_n = n^{-1/3}$



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

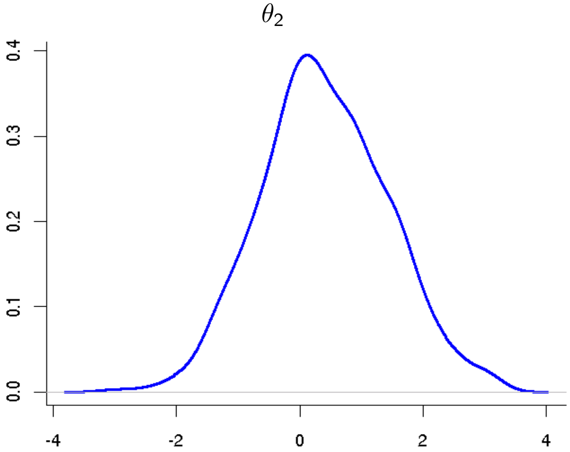
- Choose μ_n through cross-validation.



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

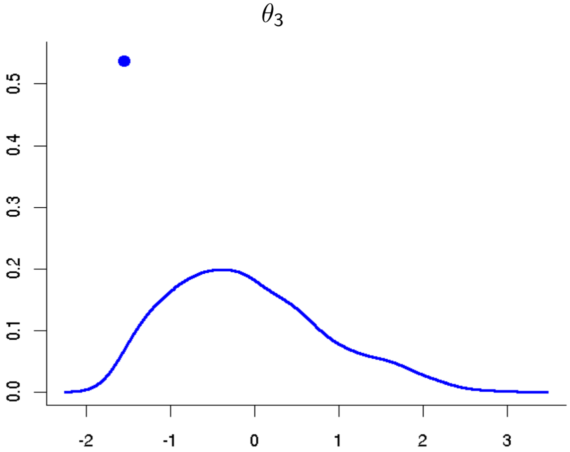
- Choose μ_n through cross-validation.



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

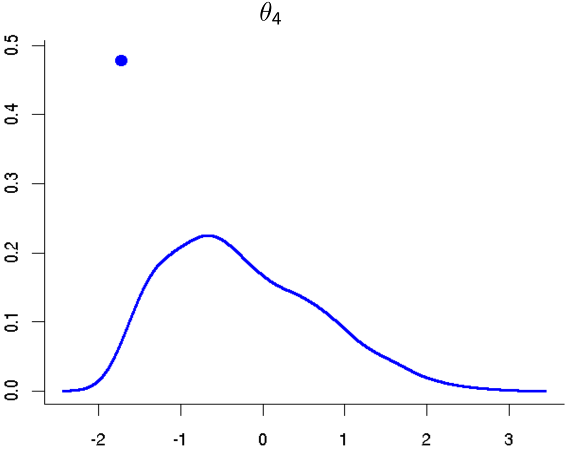
- Choose μ_n through cross-validation.



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

- Choose μ_n through cross-validation.



An impossibility result on the estimation of the cdf

Results rest on Leeb and Pötscher, 2006.

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

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

for each $\varepsilon < (\Phi(t+m) - \Phi(t-m))/2$ and each $c > 1$.

In particular no uniformly consistent estimator for $F_{n,\theta}(t)$ exists.

An impossibility result on the estimation of the cdf

Results rest on Leeb and Pötscher, 2006.

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

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

for each $\varepsilon < (\Phi(t + n^{1/2}\mu_n) - \Phi(t - n^{1/2}\mu_n))/2$, for each $c > |t|$, and for each fixed sample size n.

This is a finite-sample result for *each* estimator of $F_{n,\theta}(t)$.

Conclusions

- The finite-sample distribution of the adaptive LASSO estimator and other PMLEs 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. Relying on fixed-parameter asymptotics in this context is dangerous.
- Confidence intervals in the consistent case are larger by one order of magnitude compared to unpenalized estimator.
- Sparsity at all costs?

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.

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.

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. *Manuscript*, 2008. arXiv:0801.4627.

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.