Distributional Results for Thresholding Estimators in High-Dimensional Gaussian Regression

Ulrike Schneider

University of Göttingen

Workshop on High-Dimensional Problems in Statistics
ETH Zürich
September 23, 2011

Joint work with Benedikt Pötscher (University of Vienna)

Penalized LS (ML) Estimators

Linear regression model

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

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

A penalized least-squares (PLSE) or maximum-likelihood estimator (PMLE) $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$ is given by

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

where $X = [\mathbf{x}_{.1}, \ldots, \mathbf{x}_{.k}]$ is the $n \times k$ regressor matrix.

Penalized LS (ML) Estimators (cont'd)

- General class of Bridge-estimators (Frank & Friedman, 1993)

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

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

- Hard- and soft-thresholding estimators.
- SCAD estimator (Fan & Li, 2001)
- Elastic-net estimator (Zou & Hastie, 2005)
- Adaptive Lasso estimator (Zou, 2006)
- (thresholded) Lasso with refitting (Van de Geer et al, 2010; Belloni & Chernozhukov, 2011)
- MCP (Zhang, 2010)

  ⋮

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

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

which yields a minimum C_p-type procedure such as AIC and BIC. (l_γ-type penalty with "$\gamma = 0$") $\to$ 'classical' post-modelselection (PMS) estimators.

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

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

  Phenomenon is more pronounced for smaller γ.
- $\gamma = 1$ (Lasso and adaptive Lasso) as compromise between the wish to detect zeros and computational simplicity.

The PLSEs (and thresholding estimators) we treat in the following can be viewed to simultaneously perform model selection and parameter estimation.

Some terminology

- Consistent model selection

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

  Estimator is sparse or sparsely tuned.

- Conservative model selection

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

  Estimator is non-sparsely tuned.

Consistent vs. conservative model selection can in our context be driven by the (asymptotic) behavior of the tuning parameter.

Literature on distributional properties of PLSEs

- fixed-parameter asymptotic framework (non-uniformity issues)
- sparsely-tuned PLSEs

Oracle property – obtain same asymptotic distribution as 'oracle estimator' (infeasible unpenalized estimator using the true zero restrictions).

- Fan & Li, 2001. (SCAD)
- Zou, 2006. (Lasso and adaptive Lasso)
- Cai, Fan, Li & Zhou (2002), Fan & Li (2002, 2004), Bunea (2004), Fan & Peng (2006), Bunea & McKeague (2005), Hunter & Li (2005), Fan, Li & Zhou (2006), Wang & Leng (2007), Wang, G. Li, & Tsai (2007), Zhang & Lu (2007), Wang, R. Li, & Tsai (2007), Huang, Horowitz & Ma (2008), Li & Liang (2008), Zou & Yuan (2008), Zou & Li (2008), Johnson, Lin, & Zeng (2008), Zou & Li (2008), Zou & Yuan (2008), Lin, Xiang & Zhang (2009), Xie & Huang (2009), Zhu & Zhu (2009), Zou & Zhang (2009) ...

Literature on distributional properties of PLSEs (cont'd)

- moving-parameter asymptotic framework (taking non-uniformity into account)
- Sparsely and non-sparsely tuned PLSEs.

- Knight & Fu, 2000. (Non-sparsely tuned Lasso and Bridge estimators for $q < 1$ in general.)
- Pötscher & Leeb (2009), Pötscher & S., (2009), Pötscher & S. (2010), Pötscher & S. (2011).

Assumptions and Notation

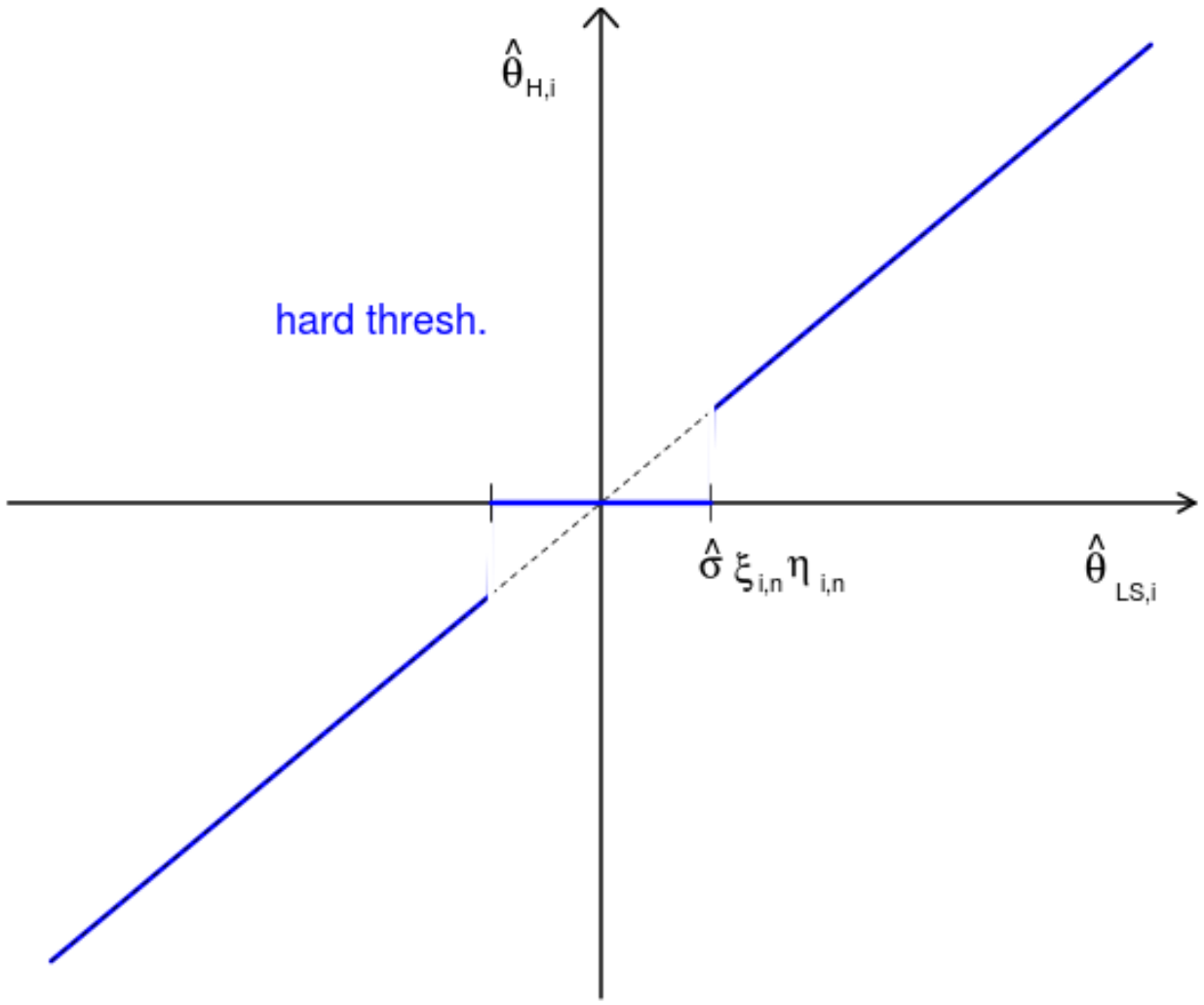
$$\mathbf{y} = X\boldsymbol{\theta} + \boldsymbol{\varepsilon}$$

- X is non-stochastic $(n \times k)$, $rk(X) = k$ $(\Rightarrow k \leq n)$. No further assumptions on X.
- k may vary with n.
- $\boldsymbol{\varepsilon} \sim N_n(0, \sigma^2 \mathcal{I}_n)$
- Notation: $\xi^2_{i,n} := \left((X'X/n)^{-1}\right)_{i,i} \quad (X'X = n\mathcal{I}_k \Rightarrow \xi_{\text{i,n}} = 1)$
- $\hat{\theta}_{\text{LS}} = (X'X)^{-1}X'y$
- $\hat{\sigma}^2_{\text{LS}} = \|y - X\hat{\theta}_{\text{LS}}\|^2/(n-k)$

Consider 3 estimators: hard-, soft- and adaptive soft-thresholding acting componentwise.

Hard-thresholding $\tilde{\theta}_{H,i}$

$$\tilde{\theta}_{H,i} = \hat{\theta}_{LS,i}\, \mathbf{1}(|\hat{\theta}_{LS,i}| > \hat{\sigma}_{LS}\xi_{i,n}\eta_{i,n})$$



Hard-thresholding $\tilde{\theta}_{H,i}$

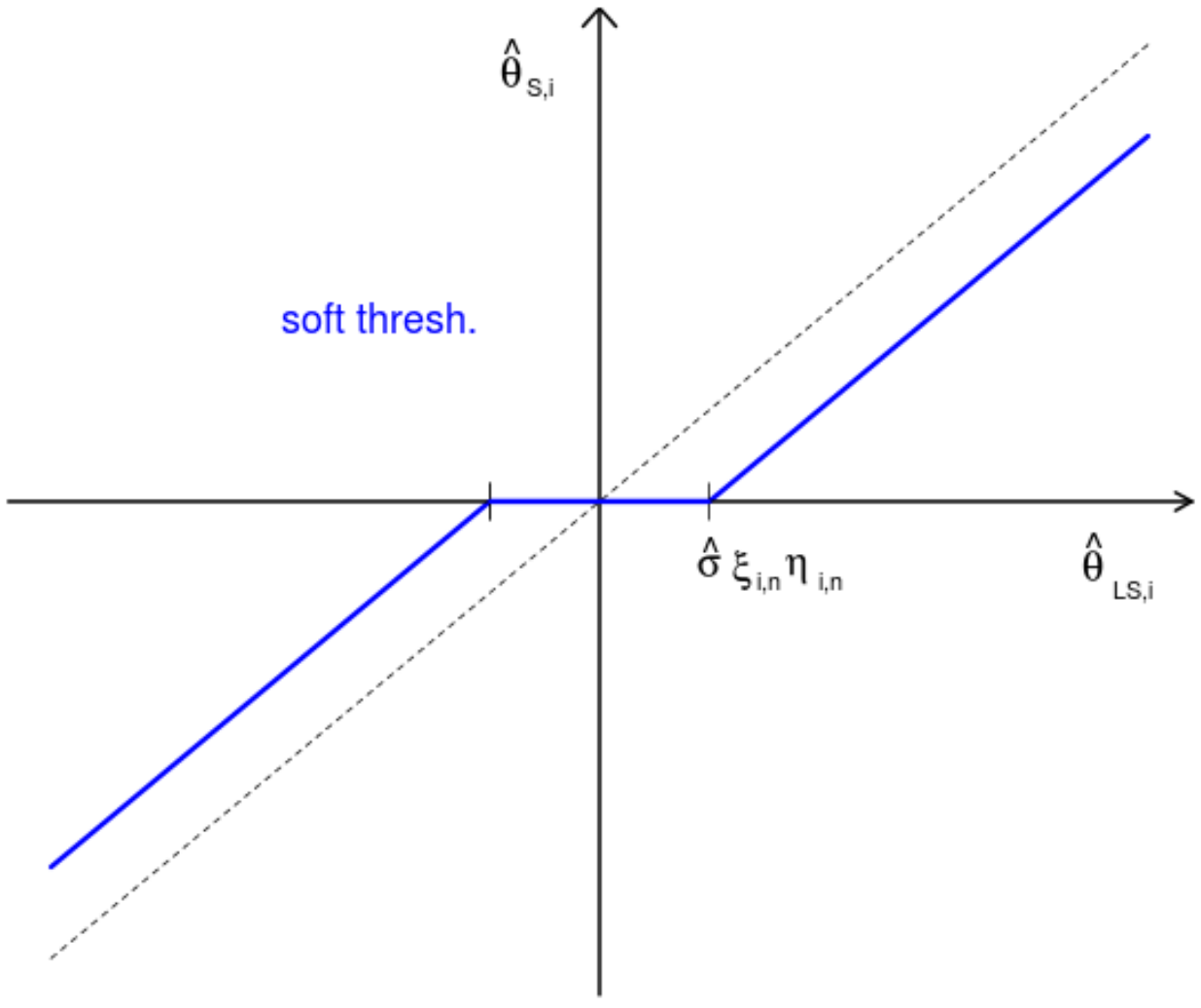
$$\tilde{\theta}_{H,i} = \hat{\theta}_{LS,i}\, \mathbf{1}(|\hat{\theta}_{LS,i}| > \hat{\sigma}_{LS}\xi_{i,n}\eta_{i,n})$$

orthogonal case:

- equivalent to a pretest estimator based on t-tests or C_p criterion such as AIC, BIC (classical post-model selection estimator) with penalty term
  $P_n(\theta) = \sum_{i=1}^{k} n \left[(\hat{\sigma}_{LS}\xi_{i,n}\eta_{i,n})^2 - (|\theta_i| - \hat{\sigma}_{LS}\xi_{i,n}\eta_{i,n})^2\, \mathbf{1}(|\theta_i| < \hat{\sigma}_{LS}\xi_{i,n}\eta_{i,n})\right]$
- also equivalent to MCP

Soft-thresholding $\tilde{\theta}_{s,i}$

$$\tilde{\theta}_{S,i} = \text{sign}(\hat{\theta}_{LS,i}) \left(|\hat{\theta}_{LS,i}| - \hat{\sigma}_{LS} \xi_{i,n} \eta_{i,n} \right)_+$$



Soft-thresholding $\tilde{\theta}_{\mathrm{S,i}}$

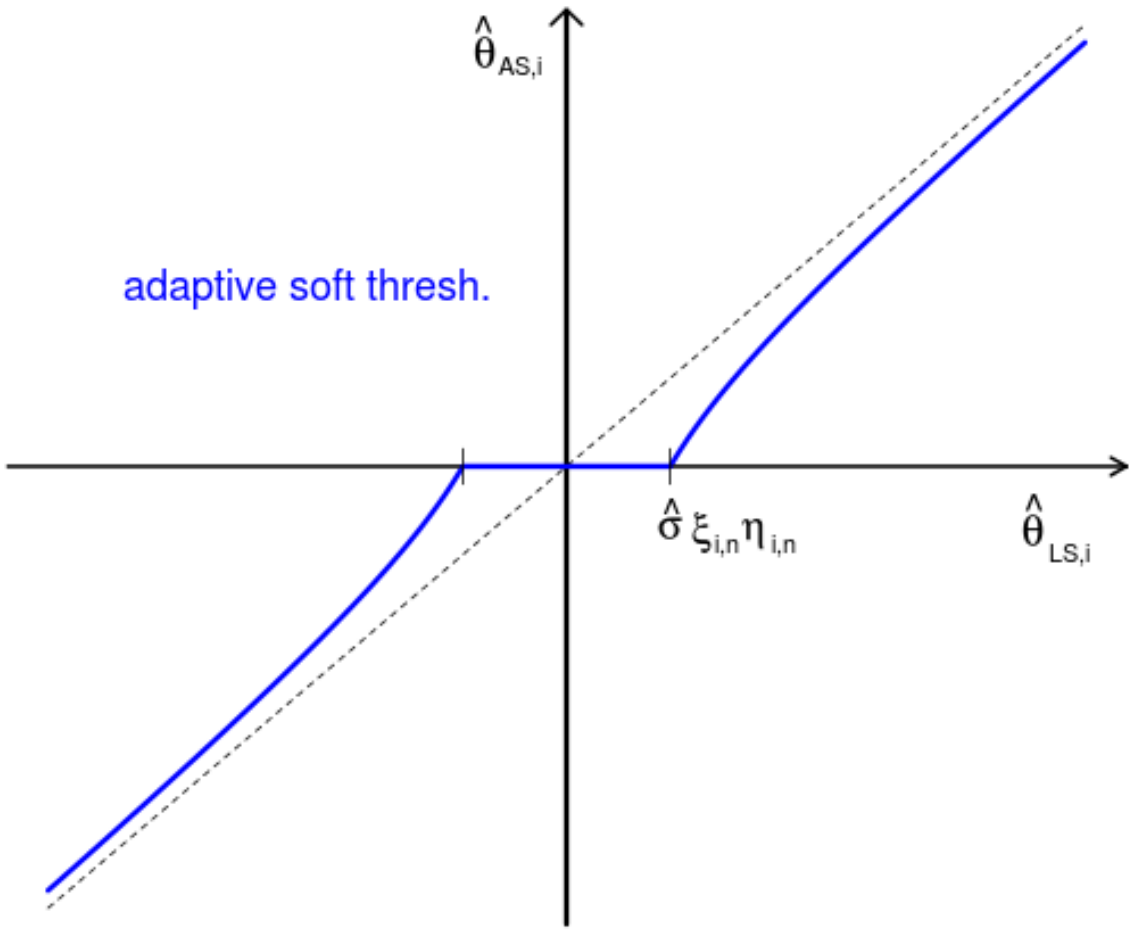
$$\tilde{\theta}_{\mathrm{S,i}} = \mathrm{sign}(\hat{\theta}_{\mathrm{LS,i}}) \left(|\hat{\theta}_{\mathrm{LS,i}}| - \hat{\sigma}_{\mathrm{LS}} \xi_{\mathrm{i,n}} \eta_{\mathrm{i,n}} \right)_+$$

orthogonal case:

- equivalent to Lasso with penalty term $P_n(\theta) = 2n\hat{\sigma}_{\mathrm{LS}} \sum_{i=1}^{k} \xi_{\mathrm{i,n}} \eta_{\mathrm{i,n}} |\theta_i|$
- also equivalent to Dantzig selector

Adaptive soft-thresholding $\tilde{\theta}_{\mathrm{AS},i}$

$$\tilde{\theta}_{\mathrm{AS},i} = \begin{cases} 0 & \text{if } |\hat{\theta}_{\mathrm{LS,i}}| \leq \hat{\sigma}_{\mathrm{LS}}\xi_{\mathrm{i,n}}\eta_{\mathrm{i,n}} \\ \hat{\theta}_{\mathrm{LS,i}} - (\hat{\sigma}_{\mathrm{LS}}\xi_{\mathrm{i,n}}\eta_{\mathrm{i,n}})^2/\hat{\theta}_{\mathrm{LS,i}} & \text{if } |\hat{\theta}_{\mathrm{LS,i}}| > \hat{\sigma}_{\mathrm{LS}}\xi_{\mathrm{i,n}}\eta_{\mathrm{i,n}} \end{cases}$$



Adaptive soft-thresholding $\tilde{\theta}_{\mathrm{AS},i}$

$$\tilde{\theta}_{\mathrm{AS},i} = \begin{cases} 0 & \text{if } |\hat{\theta}_{\mathrm{LS,i}}| \leq \hat{\sigma}_{\mathrm{LS}} \xi_{\mathrm{i,n}} \eta_{\mathrm{i,n}} \\ \hat{\theta}_{\mathrm{LS,i}} - (\hat{\sigma}_{\mathrm{LS}} \xi_{\mathrm{i,n}} \eta_{\mathrm{i,n}})^2 / \hat{\theta}_{\mathrm{LS,i}} & \text{if } |\hat{\theta}_{\mathrm{LS,i}}| > \hat{\sigma}_{\mathrm{LS}} \xi_{\mathrm{i,n}} \eta_{\mathrm{i,n}} \end{cases}$$

orthogonal case:

- equivalent to adaptive Lasso with penalty term $P_n(\theta) = 2n\hat{\sigma}^2_{\mathrm{LS}} \sum_{i=1}^{k} (\xi_{\mathrm{i,n}} \eta_{\mathrm{i,n}})^2 |\theta_i| / |\hat{\theta}_{\mathrm{LS,i}}|$
- also equivalent to non-negative Garotte (Breiman, 1995)

"Infeasible" versions

Known-variance case:

- $\hat{\theta}_{\mathsf{H,i}} = \hat{\theta}_{\mathsf{LS,i}}\, \mathbf{1}(|\hat{\theta}_{\mathsf{LS,i}}| > \sigma\xi_{\mathsf{i,n}}\eta_{\mathsf{i,n}})$
- $\hat{\theta}_{\mathsf{S,i}} = \mathsf{sign}(\hat{\theta}_{\mathsf{LS,i}})\,(|\hat{\theta}_{\mathsf{LS,i}}| - \sigma\xi_{\mathsf{i,n}}\eta_{\mathsf{i,n}})_+$
- $\hat{\theta}_{\mathsf{AS},i} = \begin{cases} 0 & \text{if } |\hat{\theta}_{\mathsf{LS,i}}| \leq \sigma\xi_{\mathsf{i,n}}\eta_{\mathsf{i,n}} \\ \hat{\theta}_{\mathsf{LS,i}} - (\sigma\xi_{\mathsf{i,n}}\eta_{\mathsf{i,n}})^2/\hat{\theta}_{\mathsf{LS,i}} & \text{if } |\hat{\theta}_{\mathsf{LS,i}}| > \sigma\xi_{\mathsf{i,n}}\eta_{\mathsf{i,n}} \end{cases}$

Variable selection

We shall assume that $\sup \xi_{\mathrm{i,n}}/n^{1/2} < \infty$.

Let $\check{\theta}_i$ stand for any of the estimators $\hat{\theta}_{\mathrm{H,i}}$, $\hat{\theta}_{\mathrm{S,i}}$, $\hat{\theta}_{\mathrm{AS},i}$, $\tilde{\theta}_{\mathrm{H,i}}$, $\tilde{\theta}_{\mathrm{S,i}}$, $\tilde{\theta}_{\mathrm{AS},i}$.

Variable selection

- $P_{n,\theta,\sigma}(\check{\theta}_i = 0) \to 0$ for any $\boldsymbol{\theta}$ with $\theta_i \neq 0 \iff \xi_{\mathrm{i,n}}\eta_{\mathrm{i,n}} \to 0$
- $P_{n,\theta,\sigma}(\check{\theta}_i = 0) \to 1$ for any $\boldsymbol{\theta}$ with $\theta_i = 0 \iff n^{1/2}\eta_{\mathrm{i,n}} \to \infty$
- $P_{n,\theta,\sigma}(\check{\theta}_i = 0) \to c_i < 1$ for any $\boldsymbol{\theta}$ with $\theta_i = 0 \iff n^{1/2}\eta_{\mathrm{i,n}} \to e_i$ with $0 \leq e_i < \infty$

1. ($\xi_{\mathrm{i,n}}\eta_{\mathrm{i,n}} \to 0$ and) $n^{1/2}\eta_{\mathrm{i,n}} \to e_i < \infty$ leads to (sensible) conservative selection.
2. ($\xi_{\mathrm{i,n}}\eta_{\mathrm{i,n}} \to 0$ and) $n^{1/2}\eta_{\mathrm{i,n}} \to \infty$ leads to (sensible) consistent selection.

Parameter estimation, minimax rate

Consistency

- $\check{\theta}_i$ is consistent for $\theta_i \iff \xi_{i,n}\eta_{i,n} \to 0$ and $\xi_{i,n}/n^{1/2} \to 0$
- Suppose $\xi_{i,n}\eta_{i,n} \to 0$ and $\xi_{i,n}/n^{1/2} \to 0$. Then $\check{\theta}_i$ is uniformly consistent for θ_i in the sense that for all $\varepsilon > 0$ there exists a real number $M > 0$ such that

$$\sup_{n\in\mathbb{N}} \sup_{\theta\in\mathbb{R}^k} \sup_{0<\sigma<\infty} P_{n,\theta,\sigma}(|\check{\theta}_i - \theta_i| > \sigma M) < \varepsilon$$

- Suppose $\xi_{i,n}\eta_{i,n} \to 0$, $\xi_{i,n}/n^{1/2} \to 0$, and $b_{i,n} \geq 0$. If for all $\varepsilon > 0$ there exists a real number $M > 0$ such that

$$\sup_{n\in\mathbb{N}} \sup_{\theta\in\mathbb{R}^k} \sup_{0<\sigma<\infty} P_{n,\theta,\sigma}(b_{i,n}|\check{\theta}_i - \theta_i| > \sigma M) < \varepsilon.$$

Then $b_{i,n} = O(a_{i,n})$, where $a_{i,n} = \min(n^{1/2}/\xi_{i,n}, (\xi_{i,n}\eta_{i,n})^{-1})$

Parameter estimation, minimax rate (cont'd)

Minimax rate is

1. $\xi_{i,n}/n^{1/2}$ in the conservative case, and
2. only $\xi_{i,n}\eta_{i,n} = o(\xi_{i,n}/n^{1/2})$ in the consistent case.

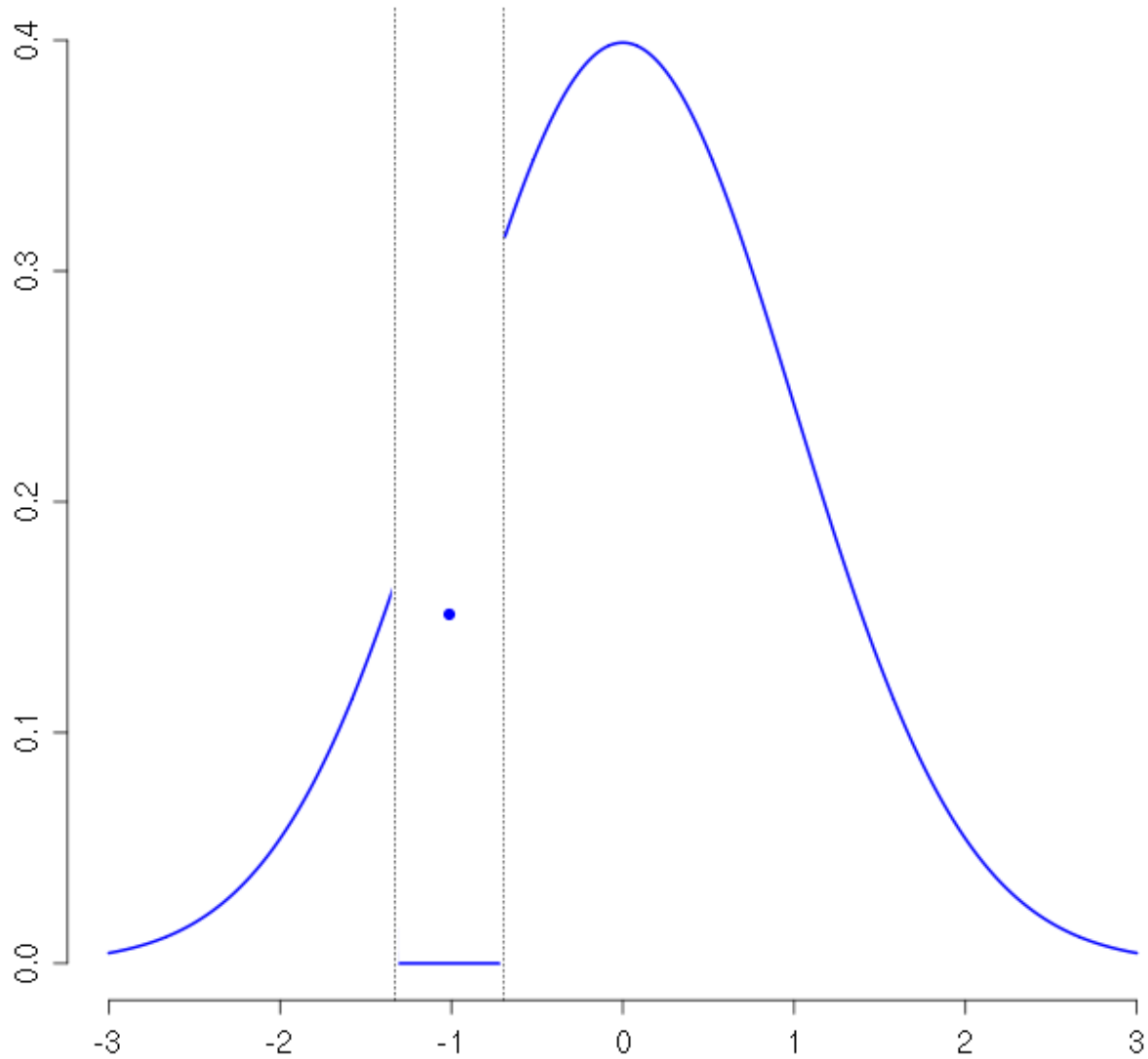
Finite sample distribution: hard-thresholding $\hat{\theta}_{\mathrm{H,i}}$

$$F^i_{\mathrm{H},n,\theta,\sigma}(x) = P_{n,\theta,\sigma}(\alpha_{\mathrm{i,n}}/\sigma(\hat{\theta}_{\mathrm{H,i}} - \theta_i) \le x) \quad \text{(known-variance case)}$$

$$\begin{aligned} dF^i_{\mathrm{H},n,\theta,\sigma}(x) \;=\; & \\ & \Big\{\Phi(n^{1/2}(-\theta_i/(\sigma\xi_{\mathrm{i,n}}) + \eta_{\mathrm{i,n}})) - \Phi(n^{1/2}(-\theta_i/(\sigma\xi_{\mathrm{i,n}}) - \eta_{\mathrm{i,n}}))\Big\}\, d\delta_{-\alpha_{\mathrm{i,n}}\theta_i/\sigma}(x) \\ & + n^{1/2}/(\alpha_{\mathrm{i,n}}\xi_{\mathrm{i,n}})\, \phi(n^{1/2}x/(\alpha_{\mathrm{i,n}}\xi_{\mathrm{i,n}}))\, \mathbf{1}(|\alpha_{\mathrm{i,n}}^{-1}x + \theta_i/\sigma| > \xi_{\mathrm{i,n}}\eta_{\mathrm{i,n}})\, dx, \end{aligned}$$

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

Finite sample distribution: hard-thresholding $\hat{\theta}_{H,i}$



$n = 40, \eta_{i,n} = 0.05, \theta_i = 0.16, \xi_{i,n} = 1, \sigma = 1, \alpha_{i,n} = n^{1/2}/\xi_{i,n}$

Finite sample distribution: hard-thresholding $\tilde{\theta}_{\mathrm{H,i}}$

$$\tilde{F}^{i}_{\mathrm{H},n,\theta,\sigma}(x) = P_{n,\theta,\sigma}(\alpha_{\mathrm{i,n}}/\sigma(\tilde{\theta}_{\mathrm{H,i}} - \theta_i) \leq x) \quad \text{(unknown-variance case)}$$

$$d\tilde{F}^{i}_{\mathrm{H},n,\theta,\sigma}(x) =$$

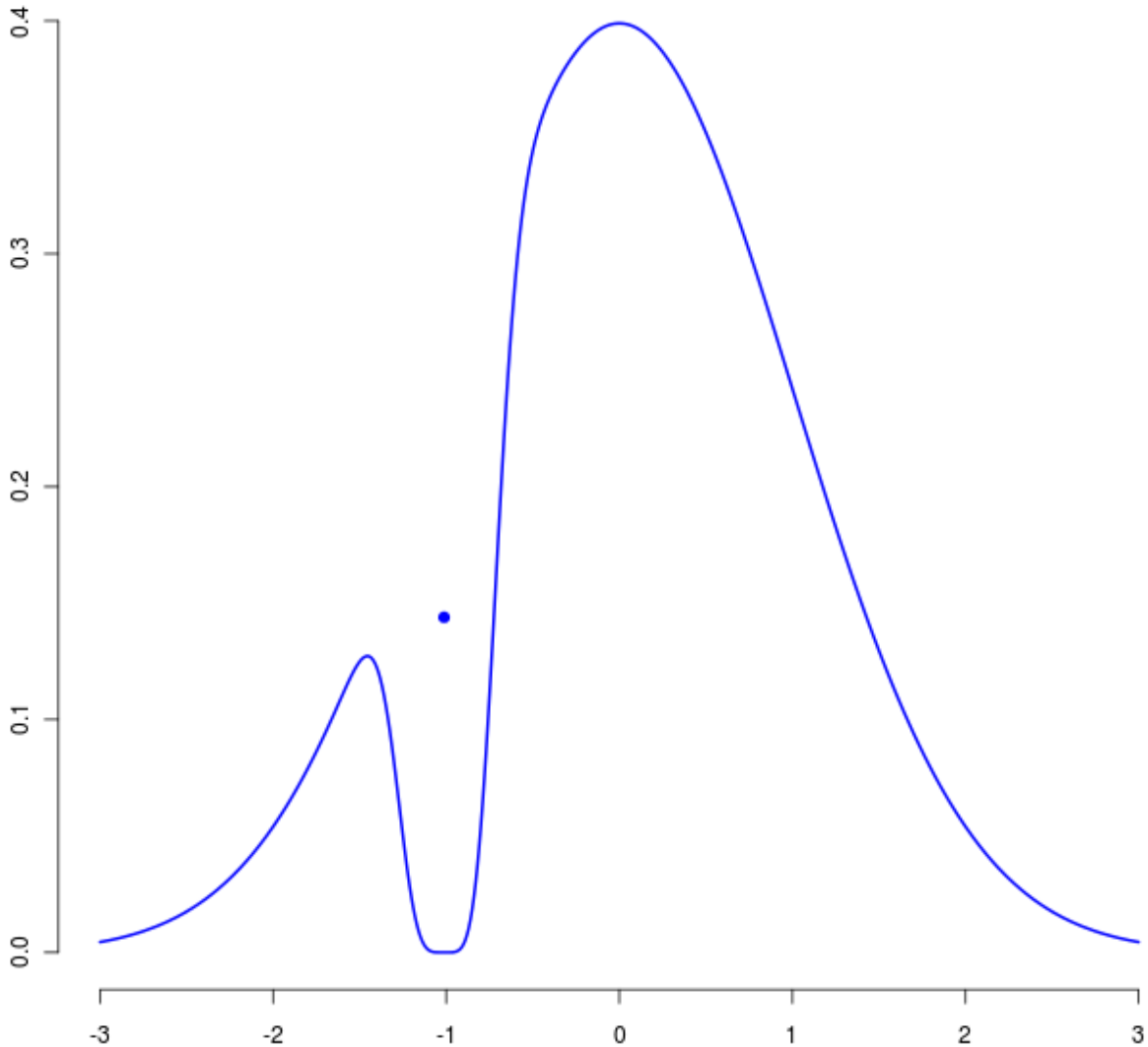
$$\int_0^\infty \{\Phi(n^{1/2}(-\theta_i/(\sigma\xi_{\mathrm{i,n}}) + s\eta_{\mathrm{i,n}}) - \Phi(n^{1/2}(-\theta_i/(\sigma\xi_{\mathrm{i,n}}) - s\eta_{\mathrm{i,n}})\}\rho_{n-k}(s)\, ds\, d\delta_{-\alpha_{\mathrm{i,n}}\theta_i/\sigma}(x)$$

$$+ n^{1/2}\alpha_{\mathrm{i,n}}^{-1}\xi_{\mathrm{i,n}}^{-1}\phi(n^{1/2}x/(\alpha_{\mathrm{i,n}}\xi_{\mathrm{i,n}})) \int_0^\infty \mathbf{1}(|\alpha_{\mathrm{i,n}}^{-1}x + \theta_i/\sigma| > \xi_{\mathrm{i,n}}s\eta_{\mathrm{i,n}})\rho_{n-k}(s)\, ds\, dx,$$

where ρ_{n-k} is the density of $\sqrt{\chi^2_{n-k}/(n-k)}$.

Finite sample distribution: hard-thresholding $\tilde{\theta}_{H,i}$

.



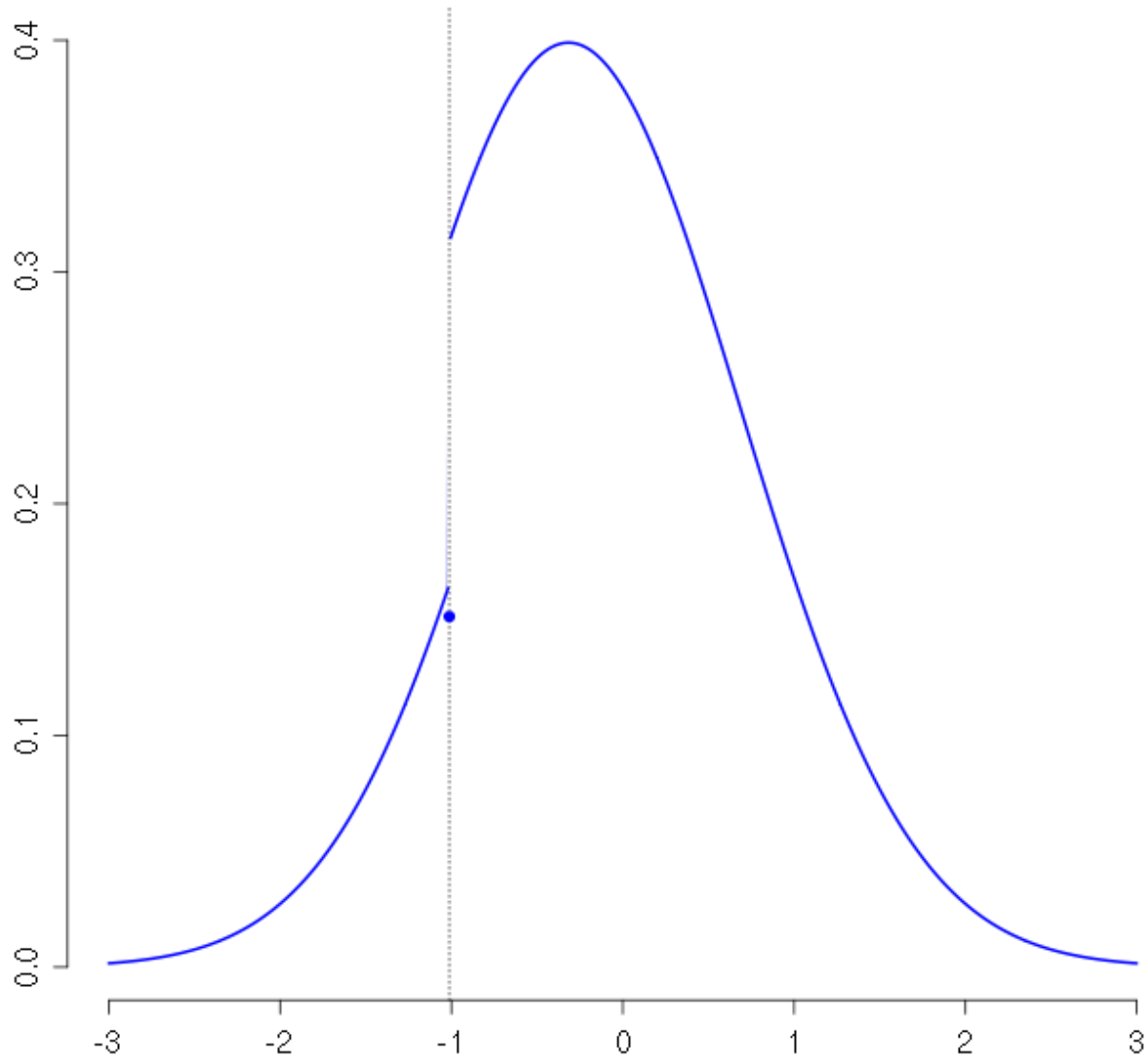
$n = 40, k = 35, \eta_{i,n} = 0.05, \theta_i = 0.16, \xi_{i,n} = 1, \sigma = 1, \alpha_{i,n} = n^{1/2}/\xi_{i,n}$

Finite sample distribution: soft-thresholding $\hat{\theta}_{S,i}$

$F^i_{S,n,\theta,\sigma}(x) = P_{n,\theta,\sigma}(\alpha_{i,n}/\sigma(\hat{\theta}_{S,i} - \theta_i) \leq x)$ (known-variance case)

$$
\begin{aligned}
dF^i_{S,n,\theta,\sigma}(x) \; = \; & \\
& \left\{ \Phi(n^{1/2}(-\theta_i/(\sigma\xi_{i,n}) + \eta_{i,n})) - \Phi(n^{1/2}(-\theta_i/(\sigma\xi_{i,n}) - \eta_{i,n})) \right\} d\delta_{-\alpha_{i,n}\theta_i/\sigma}(x) \\
& + n^{1/2}/(\alpha_{i,n}\xi_{i,n}) \left\{ \phi(n^{1/2}/(\alpha_{i,n}\xi_{i,n})x + n^{1/2}\eta_{i,n}) \mathbf{1}(\alpha_{i,n}^{-1}x + \theta_i/\sigma > 0) \right. \\
& \left. + \phi(n^{1/2}/(\alpha_{i,n}\xi_{i,n})x - n^{1/2}\eta_{i,n}) \mathbf{1}(\alpha_{i,n}^{-1}x + \theta_i/\sigma < 0) \right\} dx
\end{aligned}
$$

Finite sample distribution: soft-thresholding $\hat{\theta}_{s,i}$



$n = 40, \eta_{i,n} = 0.05, \theta_i = 0.16, \xi_{i,n} = 1, \sigma = 1, \alpha_{i,n} = n^{1/2}/\xi_{i,n}$

Finite sample distribution: soft-thresholding $\tilde{\theta}_{S,i}$

$$\tilde{F}^{i}_{S,n,\theta,\sigma}(x) = P_{n,\theta,\sigma}(\alpha_{i,n}(\tilde{\theta}_{S,i} - \theta_i) \leq x) \quad \text{(unknown-variance case)}$$

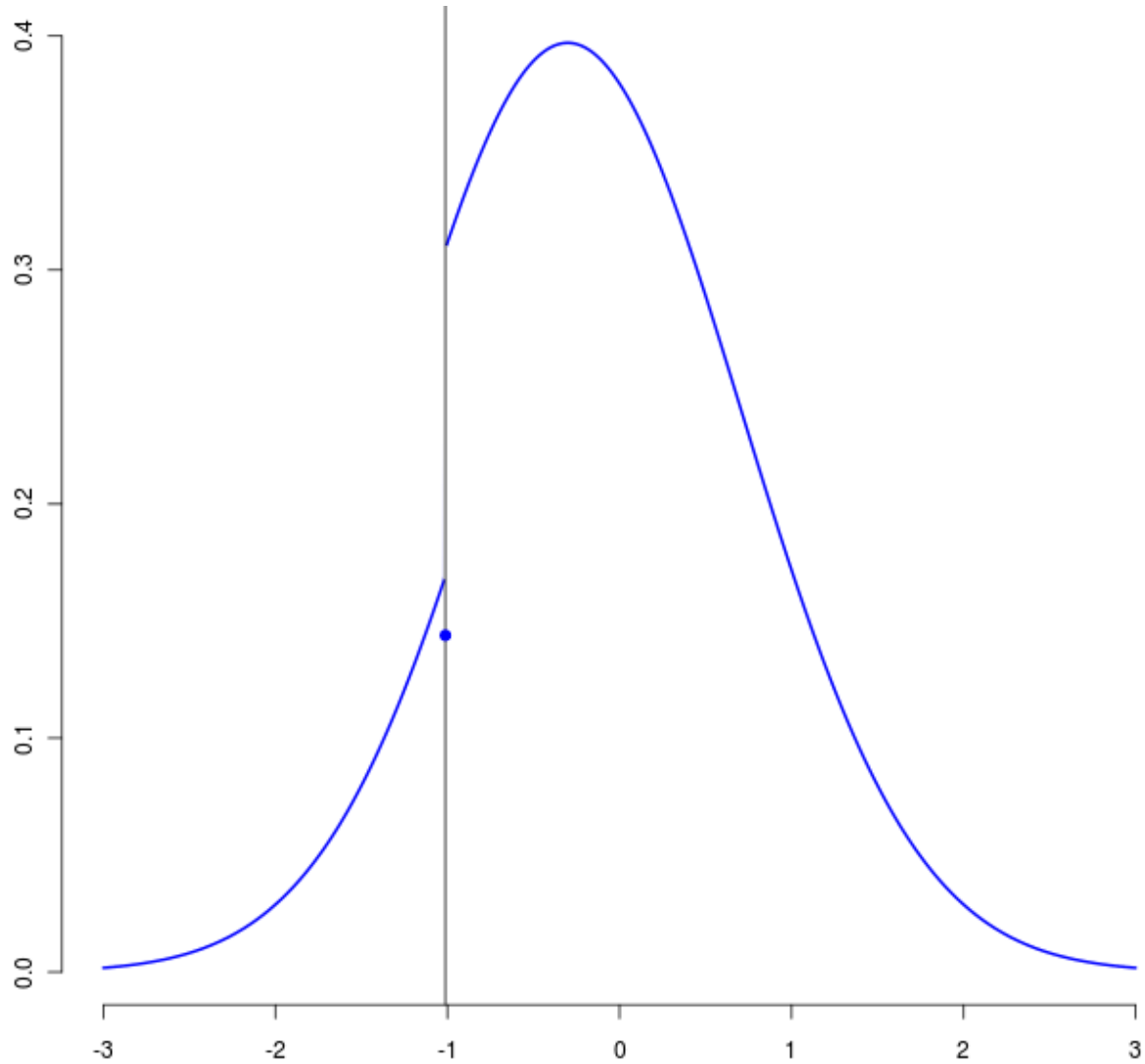
$$d\tilde{F}^{i}_{S,n,\theta,\sigma}(x) =$$

$$\int_0^\infty \{\Phi(n^{1/2}(-\theta_i/(\sigma\xi_{i,n}) + s\eta_{i,n}) - \Phi(n^{1/2}(-\theta_i/(\sigma\xi_{i,n}) - s\eta_{i,n})\}\rho_{n-k}(s)\, ds\, d\delta_{-\alpha_{i,n}\theta_i/\sigma}(x)$$

$$+ n^{1/2}/(\alpha_{i,n}\xi_{i,n}) \left\{ \int_0^\infty \phi(n^{1/2}/(\alpha_{i,n}\xi_{i,n})x + n^{1/2}s\eta_{i,n})\mathbf{1}(\alpha_{i,n}^{-1}x + \theta_i/\sigma > 0) \right.$$

$$\left. + \phi(n^{1/2}/(\alpha_{i,n}\xi_{i,n})x - n^{1/2}s\eta_{i,n})\mathbf{1}(\alpha_{i,n}^{-1}x + \theta_i/\sigma < 0) \right\} \rho_{n-k}(s)\, ds\, dx$$

Finite sample distribution: soft-thresholding $\tilde{\theta}_{s,i}$



$n = 40, k = 35, \eta_{i,n} = 0.05, \theta_i = 0.16, \xi_{i,n} = 1, \sigma = 1, \alpha_{i,n} = n^{1/2}/\xi_{i,n}$

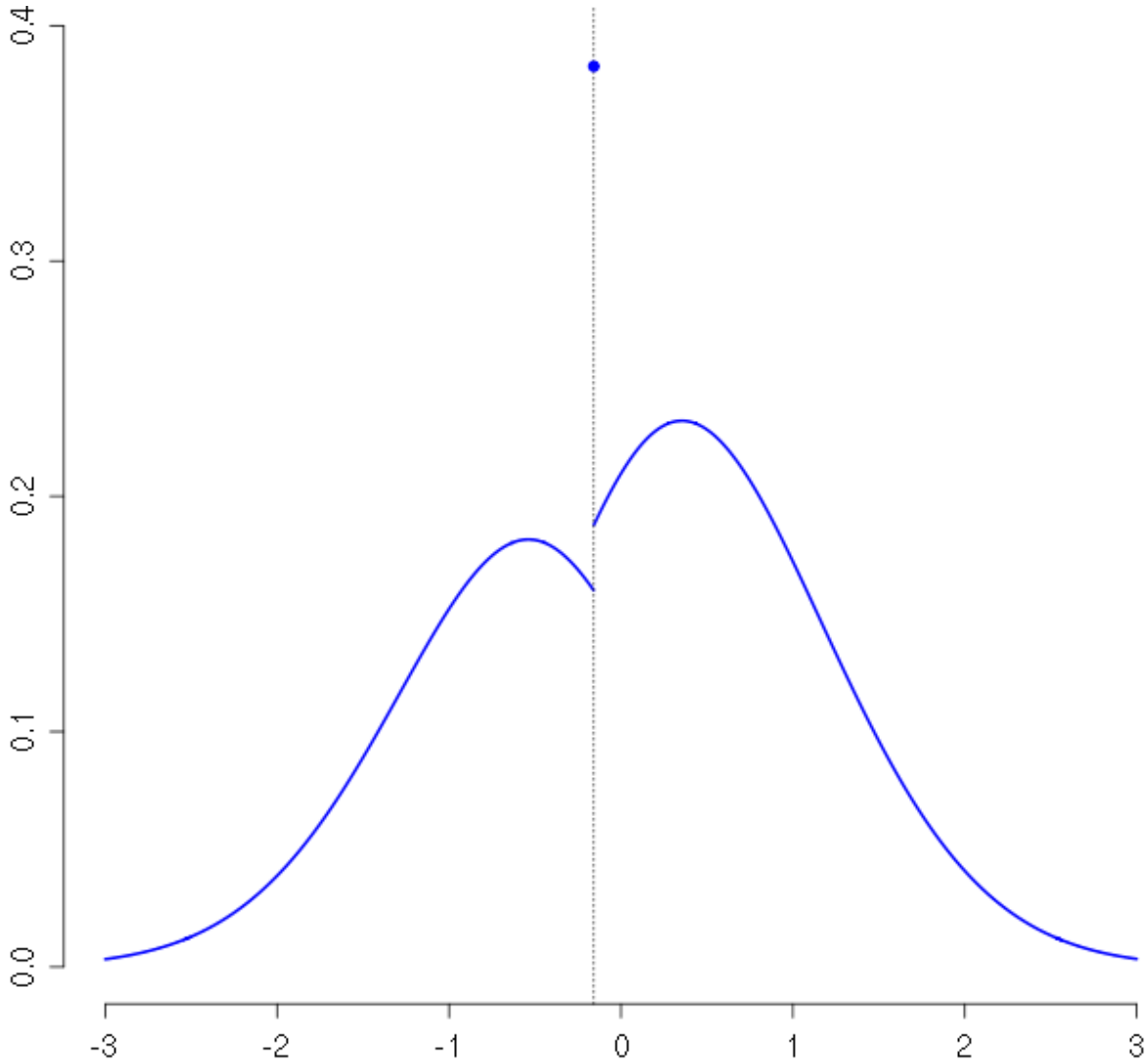
Finite sample distribution: adaptive soft-thresholding $\hat{\theta}_{\mathrm{AS},i}$

$F^i_{\mathrm{AS},n,\theta,\sigma}(x) = P_{n,\theta,\sigma}(\alpha_{\mathrm{i,n}}/\sigma(\hat{\theta}_{\mathrm{AS},i} - \theta_i) \leq x)$ (known-variance case)

$$
\begin{aligned}
&dF^i_{\mathrm{AS},n,\theta,\sigma}(x) \; = \\
&\left\{\Phi(n^{1/2}(-\theta_i/(\sigma\xi_{\mathrm{i,n}}) + \eta_{\mathrm{i,n}})) - \Phi(n^{1/2}(-\theta_i/(\sigma\xi_{\mathrm{i,n}}) - \eta_{\mathrm{i,n}}))\right\} d\delta_{-\alpha_{\mathrm{i,n}}\theta_i/\sigma}(x) \\
&+ (0.5n^{1/2}/(\alpha_{\mathrm{i,n}}\xi_{\mathrm{i,n}}))\left\{\phi(z^{(2)}_{n,\theta,\sigma}(x,\eta_{\mathrm{i,n}}))(1 + t_{n,\theta,\sigma}(x,\eta_{\mathrm{i,n}}))\mathbf{1}(\alpha^{-1}_{\mathrm{i,n}}x + \theta_i/\sigma > 0)\right. \\
&\left.+ \phi(z^{(1)}_{n,\theta,\sigma}(x,\eta_{\mathrm{i,n}}))(1 - t_{n,\theta,\sigma}(x,\eta_{\mathrm{i,n}}))\mathbf{1}(\alpha^{-1}_{\mathrm{i,n}}x + \theta_i/\sigma < 0)\right\} dx,
\end{aligned}
$$

where $z^{(1,2)}_{n,\theta,\sigma}(x,y) =$
$0.5n^{1/2}\xi^{-1}_{\mathrm{i,n}}(\alpha^{-1}_{\mathrm{i,n}}x - \theta_i/\sigma) \pm n^{1/2}\sqrt{(0.5\xi^{-1}_{\mathrm{i,n}}(\alpha^{-1}_{\mathrm{i,n}}x + \theta_i/\sigma))^2 + y^2}$ and
$t_{n,\theta,\sigma}(x,y) = 0.5\xi^{-1}_{\mathrm{i,n}}(\alpha^{-1}_{\mathrm{i,n}}x + \theta_i/\sigma)/((0.5\xi^{-1}_{\mathrm{i,n}}(\alpha^{-1}_{\mathrm{i,n}}x + \theta_i/\sigma))^2 + y^2)^{1/2}$.

Finite sample distribution: adaptive soft-thresholding $\hat{\theta}_{\mathrm{AS},i}$



$n = 40, \eta_{\mathrm{i,n}} = 0.05, \theta_i = 0.16, \xi_{\mathrm{i,n}} = 1, \sigma = 1, \alpha_{\mathrm{i,n}} = n^{1/2}/\xi_{\mathrm{i,n}}$

Finite sample distribution: adaptive soft-thresholding $\tilde{\theta}_{\mathrm{AS},i}$

$$\tilde{F}^{i}_{\mathrm{AS},n,\theta,\sigma}(x) = P_{n,\theta,\sigma}(\alpha_{\mathrm{i,n}}(\tilde{\theta}_{\mathrm{AS},i} - \theta_i) \leq x) \quad \text{(unknown-variance case)}$$

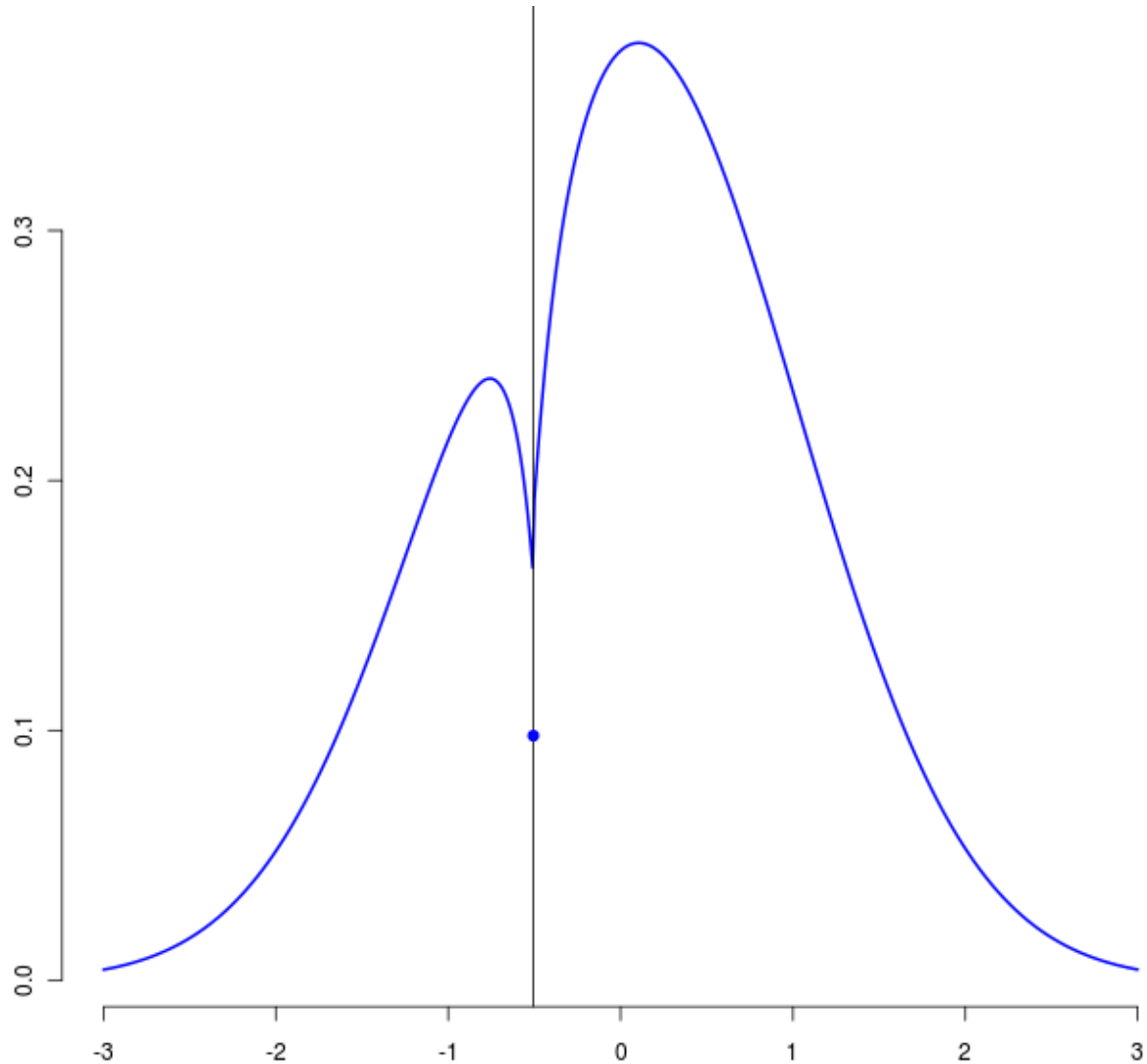
$$d\tilde{F}^{i}_{\mathrm{AS},n,\theta,\sigma}(x) \; =$$

$$\int_0^\infty \{\Phi(n^{1/2}(-\theta_i/(\sigma\xi_{\mathrm{i,n}}) + s\eta_{\mathrm{i,n}}) - \Phi(n^{1/2}(-\theta_i/(\sigma\xi_{\mathrm{i,n}}) - s\eta_{\mathrm{i,n}})\}\rho_{n-k}(s)\, ds\, d\delta_{-\alpha_{\mathrm{i,n}}\theta_i/\sigma}(x)$$

$$+ (0.5n^{1/2}/(\alpha_{\mathrm{i,n}}\xi_{\mathrm{i,n}})) \left\{ \int_0^\infty \phi(z^{(2)}_{n,\theta,\sigma}(x,\eta_{\mathrm{i,n}}))(1 + t_{n,\theta,\sigma}(x,\eta_{\mathrm{i,n}}))\mathbf{1}(\alpha_{\mathrm{i,n}}^{-1}x + \theta_i/\sigma > 0) \right.$$

$$\left. + \phi(z^{(1)}_{n,\theta,\sigma}(x,\eta_{\mathrm{i,n}}))(1 - t_{n,\theta,\sigma}(x,\eta_{\mathrm{i,n}}))\mathbf{1}(\alpha_{\mathrm{i,n}}^{-1}x + \theta_i/\sigma < 0) \right\} \rho_{n-k}(s)\, ds\, dx,$$

Finite sample distribution: adaptive soft-thresholding $\tilde{\theta}_{AS,i}$



$n = 40, k = 35, \eta_{i,n} = 0.05, \theta_i = 0.16, \xi_{i,n} = 1, \sigma = 1, \alpha_{i,n} = n^{1/2}/\xi_{i,n}$

Large sample distributions

1. Conservative tuning.

Large sample distribution: hard-thresholding $\hat{\theta}_{H,i}$

Theorem (known-variance, conservative case)

Suppose that for given $i \geq 1$ satisfying $i \leq k = k(n)$ for large enough n we have $n^{1/2}\eta_{i,n} \to e_i < \infty$. Set the scaling factor $\alpha_{i,n} = n^{1/2}/\xi_{i,n}$. Suppose that the true parameters $\theta^{(n)} = (\theta_{1,n}, \ldots, \theta_{k_n,n}) \in \mathbb{R}^{k(n)}$ and $\sigma_n \in (0, \infty)$ satisfy $n^{1/2}\theta_{i,n}/(\sigma_n \xi_{i,n}) \to \nu_i \in \mathbb{R} \cup \{-\infty, \infty\}$. Then $F^i_{H,n,\theta^{(n)},\sigma_n}$ converges weakly to the distribution with measure

$$\{\Phi(-\nu_i + e_i) - \Phi(-\nu_i - e_i)\} d\delta_{-\nu_i}(x) + \phi(x)\mathbf{1}(|x + \nu_i| > e_i)\, dx.$$

[Reduces to $N(0,1)$ if $|\nu_i| = \infty$ or $e_i = 0$.]

- Analogous results for soft-thresholding and adaptive soft-thresholding.

Uniform closeness of cdfs

Let $F^i_{\cdot,n,\theta,\sigma}$ be the cdf of either (centered and scaled) $\hat{\theta}_{\mathrm{H,i}}$, $\hat{\theta}_{\mathrm{S,i}}$.
Let $\tilde{F}^i_{\cdot,n,\theta,\sigma}$ be the cdf of either (centered and scaled) $\tilde{\theta}_{\mathrm{H,i}}$, $\tilde{\theta}_{\mathrm{S,i}}$.

Uniform closeness

Suppose that for given $i \geq 1$ satisfying $i \leq k = k(n)$ for large enough n we have $n^{1/2}\eta_{\mathrm{i,n}}(n-k)^{-1/2} \to 0$ as $n \to \infty$. Then, as $n \to \infty$

$$\sup_{\theta \in \mathbb{R}^k, 0<\sigma<\infty} \| F^i_{\cdot,n,\theta,\sigma} - \tilde{F}^i_{\cdot,n,\theta,\sigma} \|_{TV} \to 0$$

Result also holds for ad. soft-thresholding with sup-norm instead of TV-norm.

Note: If $n^{1/2}\eta_{\mathrm{i,n}} \to e_i < \infty$ (conservative case) and $n-k \to \infty$, then $n^{1/2}\eta_{\mathrm{i,n}}(n-k)^{-1/2} \to 0$ automatically holds.

Large sample distribution: hard-thresholding $\tilde{\theta}_{\mathrm{H,i}}$

Theorem (unknown-variance, conservative case)

Suppose that for given $i \geq 1$ satisfying $i \leq k = k(n)$ for large enough n we have $n^{1/2}\eta_{\mathrm{i,n}} \to e_i < \infty$. Set the scaling factor $\alpha_{\mathrm{i,n}} = n^{1/2}/\xi_{\mathrm{i,n}}$. Suppose that the true parameters $\theta^{(n)} = (\theta_{1,n}, \ldots, \theta_{k_n,n}) \in \mathbb{R}^{k(n)}$ and $\sigma_n \in (0, \infty)$ satisfy $n^{1/2}\theta_{\mathrm{i,n}}/(\sigma_n \xi_{\mathrm{i,n}}) \to \nu_i \in \mathbb{R} \cup \{-\infty, \infty\}$. Further assume that $n - k$ is eventually constant to m. Then $\tilde{F}^i_{\mathrm{H},n,\theta^{(n)},\sigma_n}$ converges weakly to the distribution with measure

$$\int_0^\infty \{\Phi(-\nu_i + se_i) - \Phi(-\nu_i - se_i)\rho_m(s)\, ds\, d\delta_{-\nu_i}(x) + \phi(x) \int_0^\infty \mathbf{1}(|x + \nu_i| > se_i)\, \rho_m(s)\, ds\, dx.$$

[Reduces to $N(0,1)$ if $|\nu_i| = \infty$ or $e_i = 0$.]

- Analogous results for soft-thresholding and adaptive soft-thresholding.

Large sample distributions

1. Conservative tuning: Asymptotic distributions capture behaviour of finite-sample distribution
   - in known variance case and
   - in the unknown variance case if $n - k$ does not diverge.

Large sample distributions

2. Consistent tuning.

Large sample distribution: hard-thresholding $\hat{\theta}_{H,i}$

Theorem (known-variance, consistent case)

Suppose that for given $i \geq 1$ satisfying $i \leq k = k(n)$ for large enough n we have $n^{1/2}\eta_{i,n} \to \infty$. Set the scaling factor $\alpha_{i,n} = (\eta_{i,n}\xi_{i,n})^{-1}$. Suppose that the true parameters $\theta^{(n)} = (\theta_{1,n}, \ldots, \theta_{k_n,n}) \in \mathbb{R}^{k(n)}$ and $\sigma_n \in (0, \infty)$ satisfy $\theta_{i,n}/(\sigma_n \xi_{i,n}\eta_{i,n}) \to \zeta_i \in \mathbb{R} \cup \{-\infty, \infty\}$. Then $F^i_{H,n,\theta^{(n)},\sigma_n}$ converges weakly to $\delta_{-\zeta_i}$ if $|\zeta_i| < 1$, and to δ_0 if $|\zeta_i| > 1$. If $|\zeta_i| = 1$, and $n^{1/2}(\eta_{i,n} - \zeta_i\theta_{i,n}/(\sigma_n\xi_{i,n})) \to r_i$, for some $r_i \in \mathbb{R}$ then the limit is $\Phi(r_i)\delta_{-\zeta_i} + (1 - \Phi(r_i))\delta_0$.

- Analogous results for soft-thresholding and adaptive soft-thresholding, except there the distributions collapse to a single pointmass in all cases.

Large sample distribution: hard-thresholding $\tilde{\theta}_{H,i}$

Theorem (unknown-variance, consistent case)

Suppose that for given $i \geq 1$ satisfying $i \leq k = k(n)$ for large enough n we have $n^{1/2}\eta_{i,n} \to \infty$. Set the scaling factor $\alpha_{i,n} = (\eta_{i,n}\xi_{i,n})^{-1}$. Suppose that the true parameters $\theta^{(n)} = (\theta_{1,n}, \ldots, \theta_{k_n,n}) \in \mathbb{R}^{k(n)}$ and $\sigma_n \in (0, \infty)$ satisfy $\theta_{i,n}/(\sigma_n \xi_{i,n}\eta_{i,n}) \to \zeta_i \in \mathbb{R} \cup \{-\infty, \infty\}$. Then $\tilde{F}^i_{H,n,\theta^{(n)},\sigma_n}$ converges weakly to

$$w(\zeta_i)\delta_{-\zeta_i} + (1 - w(\zeta_i))\delta_0$$

(a) $w(\zeta_i) = \Pr(\chi^2_m > m\zeta_i^2)$ if $n - k$ is eventually constant to $m \in \mathbb{N}$.

(b) $n - k \to \infty$: $w = 1$ if $|\zeta_i| < 1$ and $w = 0$ if $|\zeta_i| > 1$.
If $|\zeta_i| = 1$ and $n^{1/2}(\eta_{i,n} - \zeta_i\theta_{i,n}/(\sigma_n\xi_{i,n})) \to r_i \in \mathbb{R} \cup \{-\infty, \infty\}$:
1. $n^{1/2}\eta_{i,n}/(n-k)^{1/2} \to 0$: $w = \Phi(r_i)$.
2. $n^{1/2}\eta_{i,n}/(n-k)^{1/2} \to 2^{1/2}d_i$ with $0 < d_i < \infty$: $w = \int_{-\infty}^{\infty} \Phi(d_i t + r_i)\phi(t)dt$.
3. $n^{1/2}\eta_{i,n}/(n-k)^{1/2} \to \infty$ and $n^{1/2}(\eta_{i,n} - \zeta_i\theta_{i,n}/(\sigma_n\xi_{i,n}))$ / $(n^{1/2}\eta_{i,n}/(n-k)^{1/2})) \to r_i' \in \mathbb{R} \cup \{-\infty, \infty\}$: $w = \Phi(r_i')$.

Large-sample distributions

- Similar results for soft- and adaptive soft-thresholding, except that an absolutely continuous part 'survives' for the case where $n - k$ is eventually constant.

2. Consistent tuning: Asymptotic distributions always collapse at pointmasse(s)
   - in the known variance case and
   - in the unknown variance case if $n - k \to \infty$.
   - In case of hard-thresholding, some randomness 'survives' (convex combination of two pointmasses, seems to be connected to non-continuity).

- (If $n^{1/2}/\xi_{\mathrm{i,n}}$-scaling is used, then certain sequences will diverge to $\pm\infty$.)

Large-sample distributions (cont'd)

- Theorems reflect that

$$\check{\theta}_i - \theta_i = \text{“BIAS”} + \text{“FLUCTUATION”},$$

  where

  - “BIAS” is $O(\xi_{i,n}\eta_{i,n})$ ($O(n^{-1/2})$ in a pointwise sense)
  - “FLUCTUATION” is $O(n^{-1/2})$.

Honest confidence sets

Revert to simpler model:

- orthogonal design $X'X = n\mathcal{I}_n$ ($\xi_{i,n} = 1$)
- known variance $\sigma = 1$ (for presentation purposes only)

Wlog, consider a Gaussian location model: $y_1, \ldots, y_n \overset{\text{iid}}{\sim} N(\theta, 1)$.
($k = 1$, $\hat{\theta}_{\text{LS}} = \bar{y}$)

Let $\hat{\theta}$ be one the estimators $\hat{\theta}_{\text{H}}$, $\hat{\theta}_{\text{L}}$ or $\hat{\theta}_{\text{AL}}$ for θ.

We call an interval of the form $C_n = [\hat{\theta} - a, \hat{\theta} + b]$ a valid or honest confidence interval based on $\hat{\theta}$ with significance level δ, if

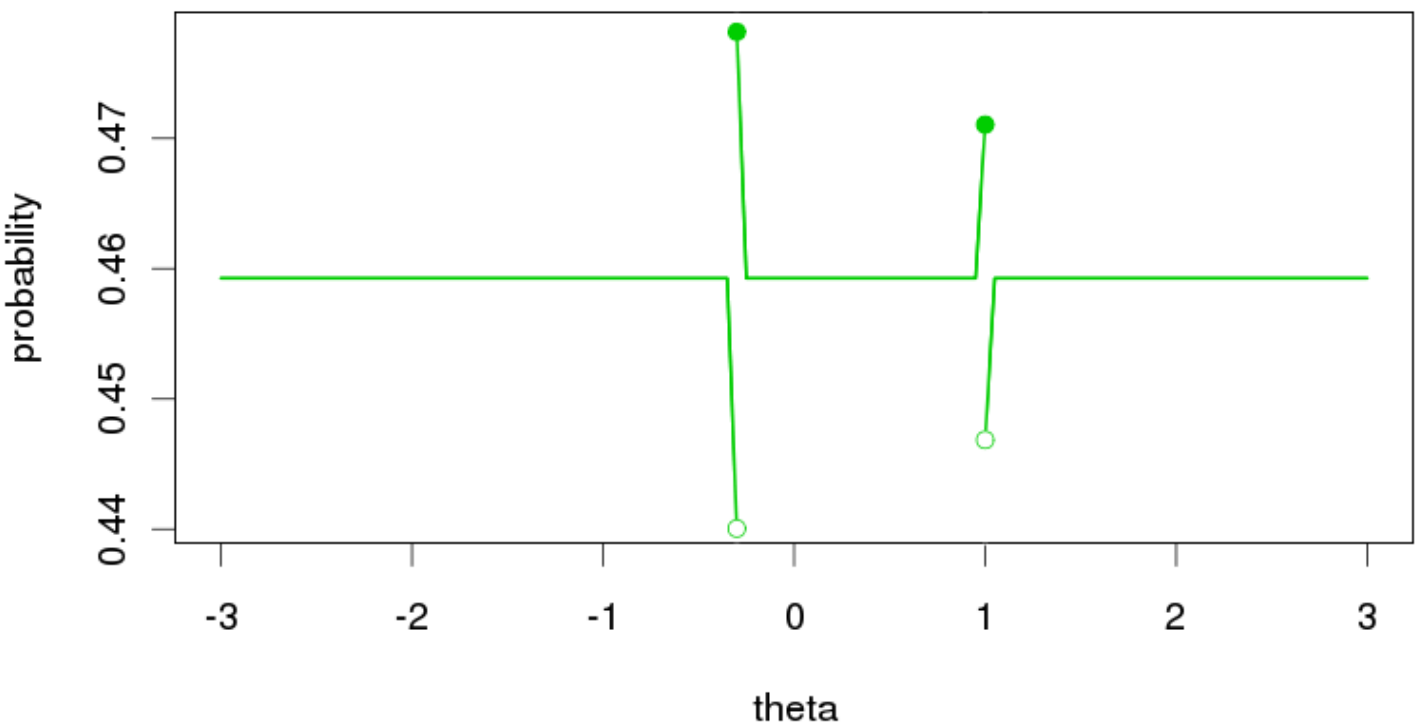
$$\inf_{\theta \in \mathbb{R}} P_{n,\theta}(\theta \in C_n) \geq \delta$$

Minimal coverage probabilities

Hard-thresholding

θ vs. $P_{n,\theta}(\theta \in C_{n,H})$, $C_{n,H} = [\hat{\theta}_{\mathrm{H}} - a_n, \hat{\theta}_{\mathrm{H}} + b_n]$

$(n = 1,\ a_n = 0.3,\ b_n = 1,\ \eta_n = 0.05)$



Minimal coverage probabilities

Theorem (Hard-thresholding)

Let $C_{n,\mathrm{H}} = [\hat{\theta}_{\mathrm{H}} - a_n, \hat{\theta}_{\mathrm{H}} + b_n]$ with $a_n, b_n \geq 0$.

$$\inf_{\theta \in \mathbb{R}} P_{n,\theta}(\theta \in C_{n,\mathrm{H}})$$

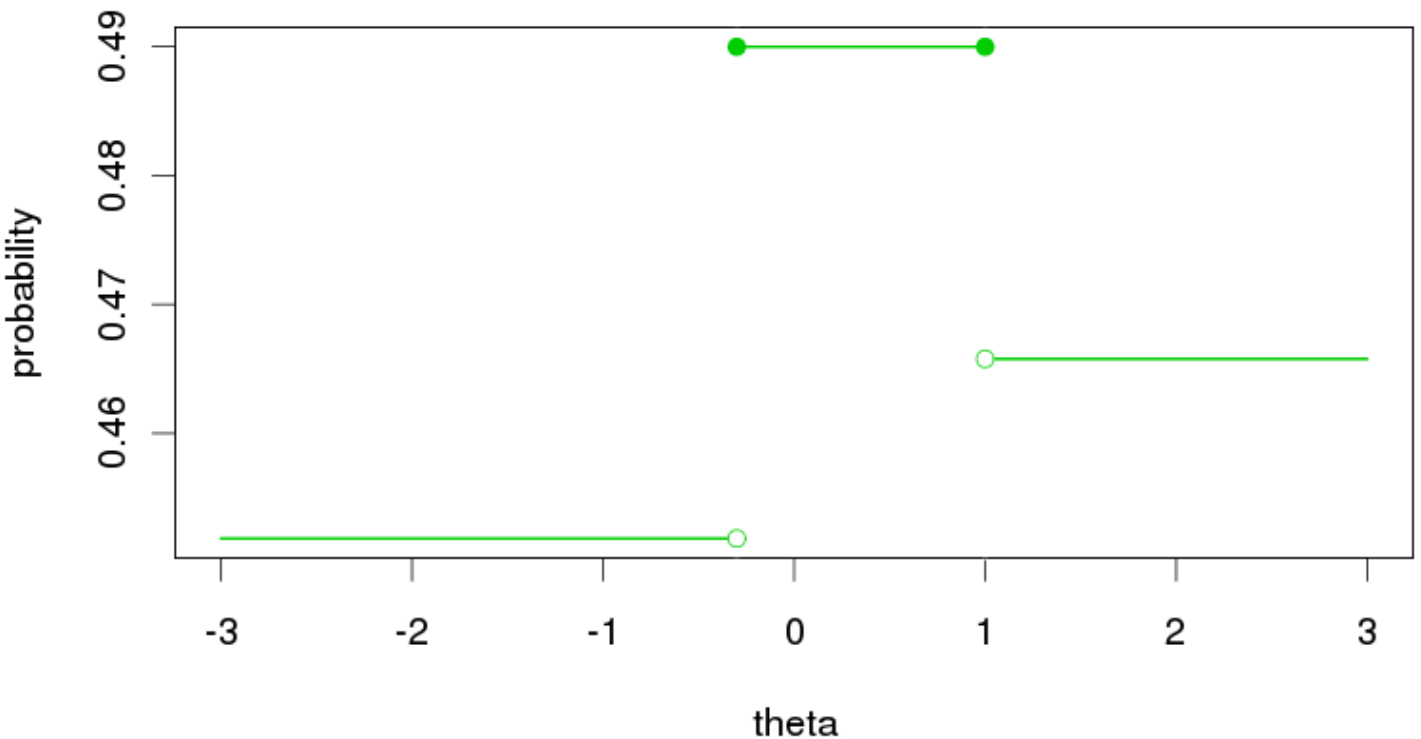
$$= \begin{cases} \Phi(n^{1/2}(a_n - \eta_n)) - \Phi(-n^{1/2} b_n) & \text{for } \eta_n \leq a_n + b_n \text{ und } a_n \leq b_n \\ \Phi(n^{1/2}(b_n - \eta_n)) - \Phi(-n^{1/2} a_n) & \text{for } \eta_n \leq a_n + b_n \text{ und } a_n > b_n \\ 0 & \text{for } \eta_n > a_n + b_n \end{cases}$$

Minimal coverage probabilities

Lasso

θ vs. $P_{n,\theta}(\theta \in C_{n,L})$, $C_{n,L} = [\hat{\theta}_{\mathrm{L}} - a_n, \hat{\theta}_{\mathrm{L}} + b_n])$

$(n = 1,\ a_n = 0.3,\ b_n = 1,\ \eta_n = 0.05)$



Minimal coverage probabilities

Theorem (Lasso)

Let $C_{n,\mathrm{L}} = [\hat{\theta}_{\mathrm{L}} - a_n, \hat{\theta}_{\mathrm{L}} + b_n]$ with $a_n, b_n \geq 0$.

$$\inf_{\theta \in \mathbb{R}} P_{n,\theta}(\theta \in C_{n,\mathrm{L}})$$

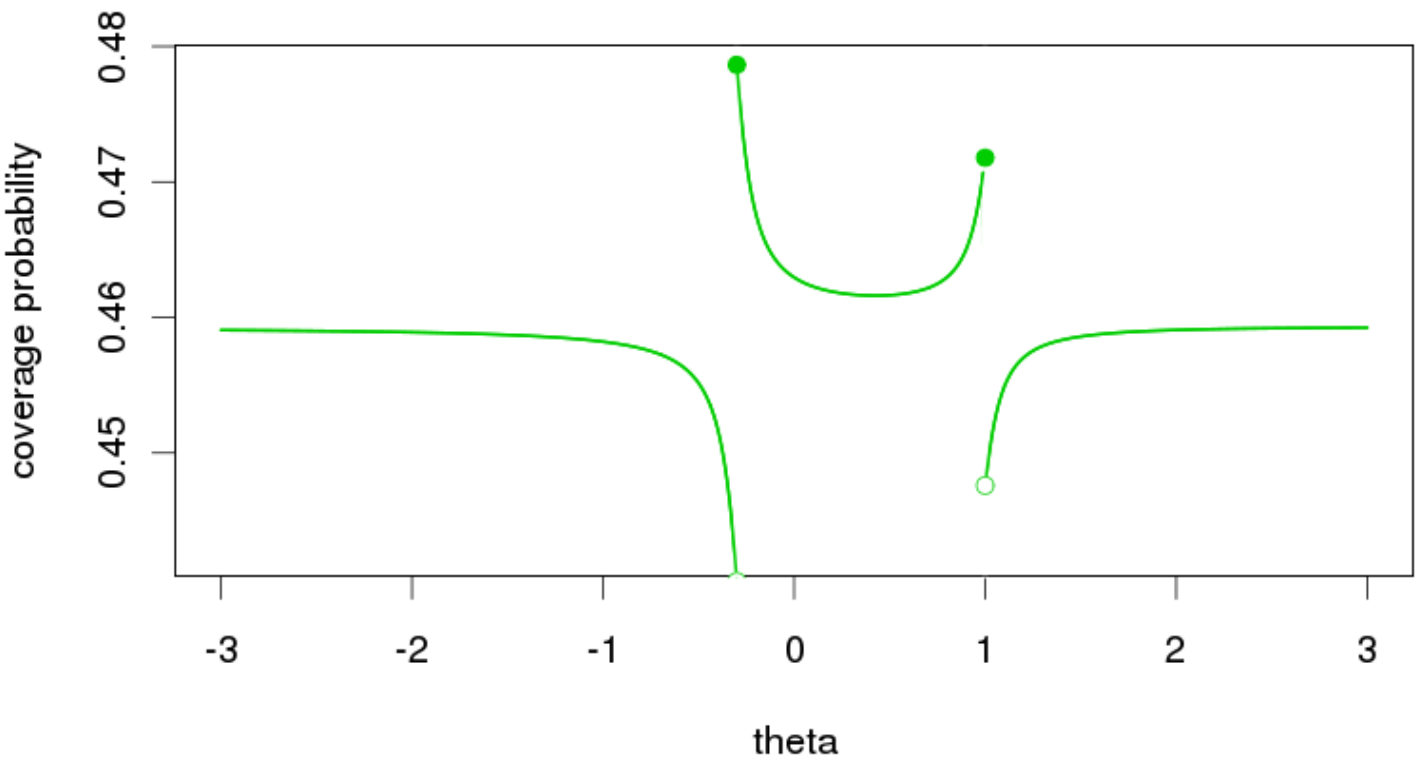
$$= \begin{cases} \Phi(n^{1/2}(a_n - \eta_n)) - \Phi(n^{1/2}(-b_n - \eta_n)) & \text{for } a_n \leq b_n \\ \Phi(n^{1/2}(b_n - \eta_n)) - \Phi(n^{1/2}(-a_n - \eta_n)) & \text{for } a_n > b_n \end{cases}$$

Minimal coverage probabilities

Adaptive Lasso

θ vs. $P_{n,\theta}(\theta \in C_{n,A})$, $C_{n,A} = [\hat{\theta}_{\text{AL}} - a_n, \hat{\theta}_{\text{AL}} + b_n])$

$(n = 1,\ a_n = 0.3,\ b_n = 1,\ \eta_n = 0.05)$



Minimal coverage probabilities

Adaptive Lasso

Let $C_{n,\mathrm{AL}} = [\hat{\theta}_{\mathrm{AL}} - a_n, \hat{\theta}_{\mathrm{L}} + b_n]$ with $a_n, b_n \geq 0$.

$$\inf_{\theta \in \mathbb{R}} P_{n,\theta}(\theta \in C_{n,\mathrm{AL}}) =$$

$$\begin{cases} \Phi\left(n^{1/2}(a_n - \eta_n)\right) - \Phi\left(n^{1/2}(a_n - b_n)/2 - \sqrt{((a_n + b_n)/2)^2 + \eta_n^2)}\right) \\ \qquad\qquad \text{for } a_n \leq b_n \\ \Phi\left(n^{1/2}((a_n - b_n)/2 + \sqrt{((a_n + b_n)/2)^2 + \eta_n^2)}\right) - \Phi\left(n^{1/2}(-b_n + \eta_n)\right) \\ \qquad\qquad \text{for } a_n > b_n \end{cases}$$

The concrete confidence intervals

Let $0 < \delta < 1$.

Hard-thresholding

Among the intervals $C_{n,\mathrm{H}}$ with minimal coverage probability not less than δ, there exists a unique shortest interval $C^*_{n,\mathrm{H}}$ with $C^*_{n,\mathrm{H}} = [\hat{\theta}_{\mathrm{H}} - a_{n,\mathrm{H}}, \hat{\theta}_{\mathrm{H}} + a_{n,\mathrm{H}}]$, where $a_{n,\mathrm{H}}$ is the unique solution of

$$\Phi(n^{1/2}(a - \eta_n)) - \Phi(-n^{1/2}a) = \delta.$$

The interval $C^*_{n,\mathrm{H}}$ has minimal coverage probability equal to δ and $a_{n,\mathrm{H}}$ is positive.

Symmetric intervals are the shortest!

The concrete confidence intervals

Let $0 < \delta < 1$.

Soft-thresholding (Lasso)

Among the intervals $C_{n,\mathrm{L}}$ with minimal coverage probability not less than δ, there exists a unique shortest interval $C^*_{n,\mathrm{L}}$ with $C^*_{n,\mathrm{L}} = [\hat{\theta}_{\mathrm{L}} - a_{n,\mathrm{L}}, \hat{\theta}_{\mathrm{L}} + a_{n,\mathrm{L}}]$, where $a_{n,\mathrm{L}}$ is the unique solution of

$$\Phi(n^{1/2}(a - \eta_n)) - \Phi(n^{1/2}(-a - \eta_n)) = \delta.$$

The interval $C^*_{n,\mathrm{L}}$ has minimal coverage probability equal to δ and $a_{n,\mathrm{L}}$ is positive.

Symmetric intervals are the shortest!

The concrete confidence intervals

Let $0 < \delta < 1$.

Adaptive Lasso

Among the intervals $C_{n,\mathrm{AL}}$ with minimal coverage probability not less than δ, there exists a unique shortest interval $C^*_{n,\mathrm{AL}}$ with $C^*_{n,\mathrm{AL}} = [\hat{\theta}_{\mathrm{AL}} - a_{n,\mathrm{AL}}, \hat{\theta}_{\mathrm{AL}} + a_{n,\mathrm{AL}}]$, where $a_{n,\mathrm{AL}}$ is the unique solution of

$$\Phi(n^{1/2}(a - \eta_n)) - \Phi(-n^{1/2}\sqrt{a^2 + \eta_n^2}) = \delta$$

The interval $C^*_{n,\mathrm{AL}}$ has minimal coverage probability equal to δ and a_{AL} is positive.

Symmetric intervals are the shortest!

Lengths of confidence sets – in finite-samples

For a fixed δ with $0 < \delta < 1$ and every $n \in \mathbb{N}$ we have

$$a_{n,\mathrm{H}} > a_{n,\mathrm{AL}} > a_{n,\mathrm{L}} > a_{n,\mathrm{LS}}.$$

Lengths of confidence sets – asymptotically

1. Conservative case.

$$a_{n,H} \sim a_{n,\mathrm{AL}} \sim a_{n,L} \sim a_{n,\mathrm{LS}} \sim n^{-1/2}$$

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

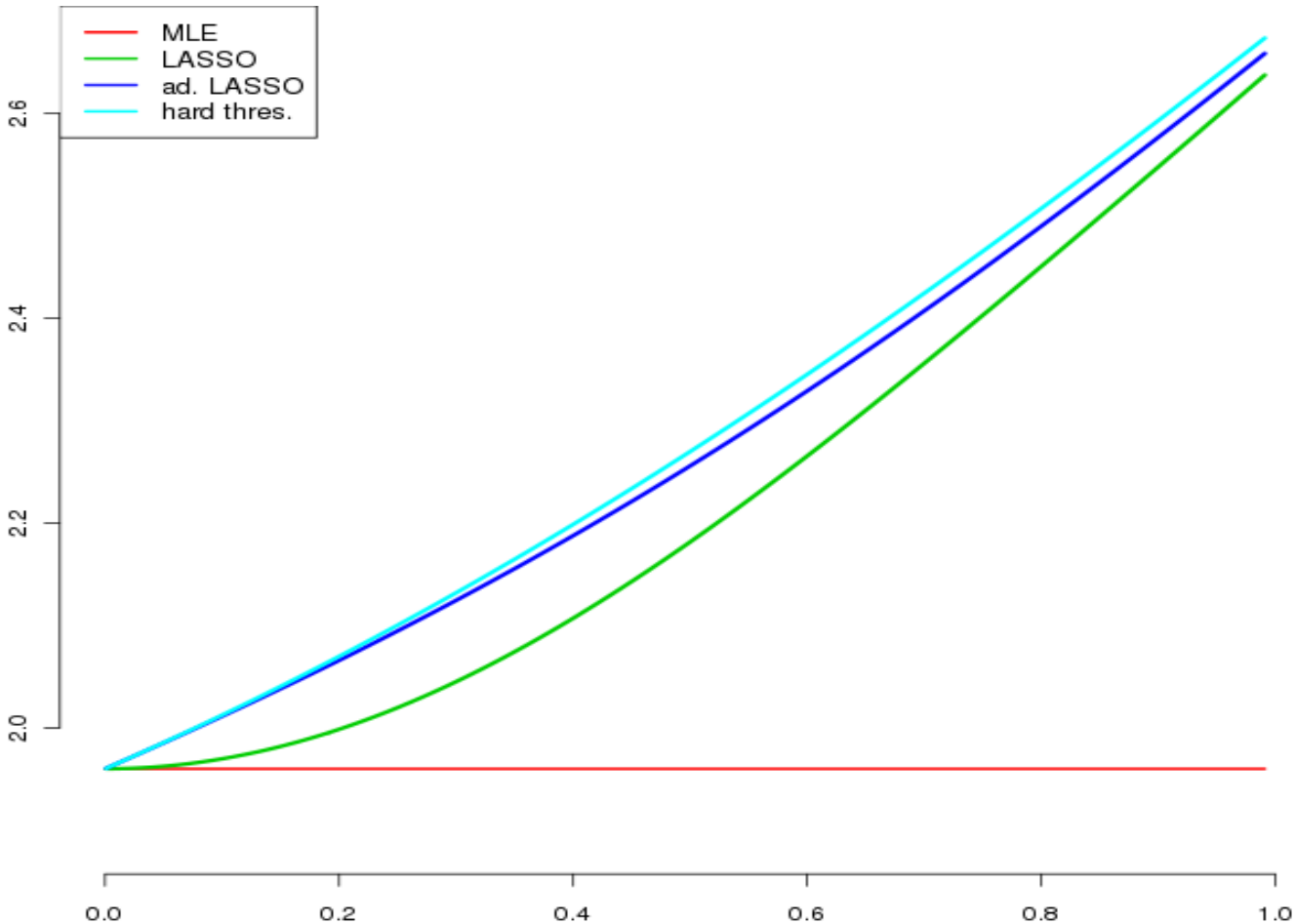
2. Consistent case. $a_{n,\{\mathrm{L,H,AL}\}} = \eta_n + n^{-1/2}\Phi(\delta) + o(n^{-1/2})$

$$a_{n,\mathrm{H}}/a_{n,\mathrm{MLE}} \sim a_{n,\mathrm{AL}}/a_{n,\mathrm{MLE}} \sim a_{n,\mathrm{L}}/a_{n,\mathrm{LS}} \sim n^{1/2}\eta_n \to \infty$$

Intervals lengths for PLSEs are larger by an order of magnitude compared to the one based the 'unpenalized' LS estimator!

Lengths of confidence sets – illustration

Plot: $n^{1/2}a_n$ vs $n^{1/2}\eta_n$ for $\delta = 0.95$.



Impossibility Results for Estimation of the cdf

Theorem

Let $\eta_n \to 0$ $n^{1/2}\eta_n \to m$ with $0 < e \leq \infty$. 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 $\varepsilon < (\Phi(t + n^{1/2}\eta_n) - \Phi(t - n^{1/2}\eta_n))/2$, for each $c > |t|$, and for each sample size n. Hence

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

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

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

Summary

We studied distributional properties of thresholding (PLS) estimators for known and unknown variance in a linear regression setting with a (potentially) growing number of parameters.

- Fixed-parameter asymptotics paint a misleading picture of the performance of the estimators.
- Finite- and large-sample distributions are highly non-normal.
- In case of consistent tuning, the uniform rate of convergence is slower than $n^{-1/2}$.
- In the unknown variance case, large-sample behavior depends on whether and how fast $n - k$ diverges in relation to the tuning parameter.
- If $n - k$ diverges, distributions collapse at point-mass for consistent tuning.

Summary (cont'd)

Orthogonal design, fixed dimension:

- Confidence sets are larger by an order of magnitude compared to the ones based on the LS-estimator in the consistent case. Lengths are of the same order for conservative tuning.

Not a criticism on the estimators per se. Distributional properties have to be investigated taking into account non-uniformity issues.

References


A. Belloni and V. Chernozhukov. Post l_1-penalized estimators in high-dimensional linear regression models. manuscript *arxiv:1001.0188*, 2010.


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.


B. M. Pötscher and H. Leeb. On the distribution of penalized maximum likelihood estimators: The LASSO, SCAD, and thresholding. *J. Multivariate Anal.*, 100:2065–2082,2009.


B. M. Pötscher and U. Schneider. On the distribution of the adaptive lasso estimator. *J. Stat. Plan. Inf.*, 139:2775–2790, 2009.


B. M. Pötscher and U. Schneider. Confidence sets based on penalized maximum likelihood estimators in Gaussian regression. *J. Electron. Stat.*, 4:334–360, 2010.


B. M. Pötscher and U. Schneider. Distributional results for thresholding estimators in high-dimensional Gaussian regression models. manuscript *arxiv:1106.6002*, 2011.


S. van de Geer and P. Bühlmann and S. Zhou. The adaptive and the thresholded Lasso for potentially misspecified models. manuscript *arxiv:1001.5176*, 2010.


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


H. Zhang. Nearly unbiased variable selection under minimax concave penalty *Ann. Stat.*, 38:894–942, 2010.


H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *J. Roy. Stat. Soc. B*, 67:301–320, 2005.