Proximal Estimation and Inference

Alberto Quaini[a] **Fabio Trojani**[b]

[a]Erasmus University
[b]University of Geneva, University of Turin and SFI

ESEM 2024, Rotterdam

Proximal Estimation

Let $\hat{\boldsymbol{\beta}}_n^s$ be an initial estimator for a target parameter of interest $\boldsymbol{\beta}_0 \in \mathbb{R}^p$

Definition (Proximal Estimator (PE))

Given symm. pos. def. matrix $\boldsymbol{W}_n$, convex lsc f_n and $\lambda_n > 0$:

$$\hat{\boldsymbol{\beta}}_n := \mathrm{prox}_{\lambda_n f_n}^{\boldsymbol{W}_n}(\hat{\boldsymbol{\beta}}_n^s) := \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2} \left\| \hat{\boldsymbol{\beta}}_n^s - \boldsymbol{\beta} \right\|_{\boldsymbol{W}_n}^2 + \lambda_n f_n(\boldsymbol{\beta}) \right\}$$

is called a proximar estimator of $\boldsymbol{\beta}_0$

- PEs are penalized minimum distance corrections defined via a (differentiable) proximal operator $\mathrm{prox}_{\lambda_n f_n}^{\boldsymbol{W}_n}$

Well-known convex lsc penalties of class $\Gamma(\mathbb{R}^p)$

- Ridge: $f_n(\boldsymbol{\beta}) = \frac{1}{2} \|\boldsymbol{\beta}\|_2^2$
- Lasso: $f_n(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1$
- Elastic Net: $f_n(\boldsymbol{\beta}) = \frac{\alpha}{2} \|\boldsymbol{\beta}\|_2^2 + (1-\alpha) \|\boldsymbol{\beta}\|_1$ for some $\alpha \in (0,1)$
- Adaptive Lasso: $f_n(\boldsymbol{\beta}) = \sum_{k=1}^{K} \frac{|\beta_i|}{|\tilde{\beta}_{ni}|}$ for some consistent auxiliary estimator $\tilde{\boldsymbol{\beta}}_n$
- Convex constraints: $f_n(\boldsymbol{\beta}) = \begin{cases} 0 & \boldsymbol{\beta} \in C \\ \infty & \boldsymbol{\beta} \notin C \end{cases}$ for some convex set C

A convenient framework

- A wide class of PEs from different choices of $\hat{\beta}_n^s$, W_n and f_n
- Embeds naturally Penalized Least Squares Estimators (PLSEs):

$$\operatorname{prox}_{\lambda_n f_n}^{\boldsymbol{X}'\boldsymbol{X}/n}(\hat{\beta}_n^{ls}) = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2} \|\boldsymbol{Y} - \boldsymbol{X}\beta\|_2^2 + \lambda_n f_n(\beta) \right\}$$

- PEs' properties derived within simple/unifying convex analysis framework:
  - Asymptotic distribution, Oracle properties,...
  - Weak/transparent assumptions on initial estimator $\hat{\beta}_n^{ls}$ and convex penalties f_n
  - Mainly determined by properties of penalties' subgradient
- PEs can be developed to handle irregular designs where PLSEs may be ill-behaved

Asymptotic properties of PEs

Main high-level assumptions

- Properties of PEs determined by triplet $(\hat{\boldsymbol{\beta}}_n^s, \boldsymbol{W}_n, \lambda_n f_n)$:

$$\hat{\boldsymbol{\beta}}_n := \text{prox}_{\lambda_n f_n}^{\boldsymbol{W}_n}(\hat{\boldsymbol{\beta}}_n^s) := \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2} \left\| \hat{\boldsymbol{\beta}}_n^s - \boldsymbol{\beta} \right\|_{\boldsymbol{W}_n}^2 + \lambda_n f_n(\boldsymbol{\beta}) \right\}$$

- Main Assumptions (Consistency, Asymptotic Distribution)

  A1 $\boldsymbol{W}_n \to_{\Pr} \boldsymbol{W}_0$ for some positive definite matrix $\boldsymbol{W}_0$

  A2 $r_n(\hat{\boldsymbol{\beta}}_n^s - \boldsymbol{\beta}_0) \to_d \boldsymbol{\eta}$, for some rate $r_n \to \infty$ and some random variable η

  A3 $r_n \lambda_n f_n \to_{\Pr} g_0$ in epigraph, for some limit penalty g_0

- Applicable to general class of convex lsc penalties and to irregular designs

Asymptotic distribution

Proposition (Asymptotic distribution of PEs)

Under Assumptions A1–A3:

$$r_n\left(\mathrm{prox}_{\lambda_n f_n}^{W_n}(\hat{\boldsymbol{\beta}}_n^s)-\boldsymbol{\beta}_0\right)\to_d \mathrm{prox}_{{g_0}'(\cdot;\boldsymbol{\beta}_0)}^{W_0}(\boldsymbol{\eta})=\left(Id-P_{\partial g_0(\boldsymbol{\beta}_0)}^{W_0}\right)(\boldsymbol{\eta})$$

with directional derivative [subgradient] ${g_0}'(\cdot;\boldsymbol{\beta}_0)$ [$\partial g_0(\boldsymbol{\beta}_0)$] of limit penalty g_0 at $\boldsymbol{\beta}_0$, and projection operator:

$$P_{\partial g_0(\boldsymbol{\beta}_0)}^{W_0}(\boldsymbol{\eta}):=\arg\min_{\boldsymbol{\theta}\in\partial g_0(\boldsymbol{\beta}_0)}\|\boldsymbol{\eta}-\boldsymbol{\theta}\|_{W_0}$$

- Functional characterization of PEs' asy. distribution via limit penalty subgradient $\partial g_0(\boldsymbol{\beta}_0)$
- Closed-form asymptotic distributions for established penalties in the literature

Example: Adaptive Lasso

- Adaptive Lasso penalty $f_n(\boldsymbol{\beta}) = \sum_{k=1}^{K} \frac{|\beta_i|}{|\tilde{\beta}^s_{ni}|}$

- If $\lambda_n r_n \to 0$ and $\lambda_n r_n^2 \to \infty$, then $\partial g_0(\boldsymbol{\beta}_0) = \text{span}\{e_j : j \in \mathcal{A}\}^\perp$:

$$r_n \left(\text{prox}^{\boldsymbol{W}_n}_{\lambda_n f_n}(\hat{\boldsymbol{\beta}}^s_n) - \boldsymbol{\beta}_0 \right) \to_d P^{\boldsymbol{W}_0}_{\text{span}\{e_j : j \in \mathcal{A}\}}(\boldsymbol{\eta})$$

- Regular linear regression model with spherical errors:
  - $\hat{\boldsymbol{\beta}}^s_n = \hat{\boldsymbol{\beta}}^{ls}_n$
  - $\boldsymbol{W}_n = \boldsymbol{X}'\boldsymbol{X}/n \to \boldsymbol{Q}_0 := \mathbb{E}[\boldsymbol{X}_1 \boldsymbol{X}_1']$ and $\boldsymbol{X}'\boldsymbol{\epsilon}/\sqrt{n} \to_d \boldsymbol{Z} \sim \mathcal{N}(\mathbf{0}, \sigma_\epsilon \boldsymbol{Q}_0)$

- Adaptive Lasso PLSE's (efficient) asy. distribution:

$$\sqrt{n} \left(\text{prox}^{\boldsymbol{Q}_n}_{\lambda_n f_n}(\hat{\boldsymbol{\beta}}^{ls}_n) - \boldsymbol{\beta}_0 \right) \to_d P^{\boldsymbol{Q}_0}_{\text{span}\{e_j : j \in \mathcal{A}\}}(\boldsymbol{Q}_0{}^{-1} \boldsymbol{Z}) = \begin{cases} [(\boldsymbol{Q}_0)_{\mathcal{A}}]^{-1} (\boldsymbol{Z})_{\mathcal{A}} & \text{in } \mathcal{A} \\ 0 & \text{in } \mathcal{A}^c \end{cases}$$

Variable selection

Definition (Variable Selection)

Given a sequence of PE's estimated active sets:

$$\hat{\mathcal{A}}_n = \left\{ j : \left(\operatorname{prox}_{\lambda_n f_n}^{W_n}(\hat{\beta}_n^s) \right)_j \neq 0 \right\} \ ,$$

$\operatorname{prox}_{\lambda_n f_n}^{W_n}(\hat{\beta}_n^s)$ is said to perform consistent variable selection (VS) if $\mathbb{P}(\hat{\mathcal{A}}_n = \mathcal{A}) \to 1$.

Proposition (Variable Selection)

Let Assumptions A1–A3 be satisfied. IF VS holds then, as $n \to \infty$:

$$\mathbb{P}\left((\eta)_{\mathcal{A}^c} = \left(P_{\partial g_0(\beta_0)}^{W_0}(\eta) \right)_{\mathcal{A}^c} \right) = 1 \ .$$

Conversely, VS holds if optimal subgradient vectors $v_n^{opt} = W_n(\hat{\beta}_n^s - \hat{\beta}_n) \in \lambda_n \partial f_n(\hat{\beta}_n)$ are such that:

$$r_n \left\| (v_n^{opt})_{\mathcal{A}^c} \right\|_1 \to_{\Pr} \infty \quad \text{as } n \to \infty$$

Example: Adaptive Lasso

- Adaptive Lasso optimal subgradient vector implied by penalty $f_n(\boldsymbol{\beta}) = \sum_{k=1}^{K} \frac{|\beta_i|}{|\tilde{\beta}_{ni}|}$ yields:

$$r_n \left\| (\boldsymbol{v}_n^{opt})_{\mathcal{A}^c} \right\|_1 = \lambda_n r_n \sum_{j \in \mathcal{A}^c} (1/|\tilde{\beta}_{nj}|)$$

- Whenever $\lambda_n r_n^2 \to \infty$ and $(\tilde{\boldsymbol{\beta}}_n)_{\mathcal{A}^c} = O_{\text{Pr}}(1/r_n)$:

$$r_n \left\| (\boldsymbol{v}_n^{opt})_{\mathcal{A}^c} \right\|_1 \to_{\text{Pr}} \infty$$

i.e., VS holds.

PEs for irregular designs

Linear regression model with irregular design

- Linear model $\boldsymbol{Y} = \boldsymbol{X}\beta_0 + \epsilon$
- Sample design matrix $\boldsymbol{Q}_n := \boldsymbol{X}'\boldsymbol{X}/n$
- Population design matrix $\boldsymbol{Q}_{0n} := \mathbb{E}[\boldsymbol{Q}_n]$

Definition (Irregular design)

(i) Singular design. There exists singular matrix $\boldsymbol{Q}_0$ such that:

$$\boldsymbol{Q}_{0n} = \boldsymbol{Q}_0, \quad \text{for all } n$$

(ii) Nearly-singular design. $\boldsymbol{Q}_{0n}$ is regular for all n and there exists singular matrix $\boldsymbol{Q}_0$ such that:

$$\boldsymbol{Q}_{0n} \to \boldsymbol{Q}_0 \text{ as } n \to \infty$$

Proximal estimation approach to irregular designs

- Under an irregular design, the set of limit population LS solutions is not a singleton
- Introduce a convenient identifiable parameter $\boldsymbol{\beta}_0 \in \mathbb{R}^p$
- Build initial estimator $\hat{\boldsymbol{\beta}}_n^s$ of $\boldsymbol{\beta}_0$, which is well-behaved under both regular and irregular designs
- Build suitable proximal estimator $\mathrm{prox}_{\lambda_n f_n}^{W_n}(\hat{\boldsymbol{\beta}}_n^s)$ of $\boldsymbol{\beta}_0$, which ideally satisfies the Oracle property

Ridgeless (limit) population parameter

- Let $\boldsymbol{\delta}_0 := \lim_{n\to\infty} \mathbb{E}[\boldsymbol{X}'\boldsymbol{Y}/n]$

Definition (Ridgeless target parameter)

Given Moore-Penrose inverse $\boldsymbol{Q}_0^+$, the Ridgeless population parameter is given by:

$$\beta_0^+ := \arg\min_{\boldsymbol{\beta}}\{\|\beta\|_2 : \boldsymbol{Q}_0\boldsymbol{\beta} = \boldsymbol{\delta}_0\} = \boldsymbol{Q}_0^+\boldsymbol{\delta}_0$$

- β_0^+ is identified under both a regular and an irregular design

Ridgeless estimator

Definition (Ridgeless estimator)

$$\hat{\beta}_n^+ := \arg\min_{\boldsymbol{\beta}} \{\|\beta\|_2 : \boldsymbol{Q}_n \beta = \boldsymbol{X}'\boldsymbol{Y}/n\} = \boldsymbol{Q}_n^+ \boldsymbol{X}'\boldsymbol{Y}/n$$

- $\hat{\beta}_n^+ = \hat{\beta}_n^{ls}$ if $\boldsymbol{Q}_n$ is regular
- Using standard assumptions, $\hat{\beta}_n^+$ is $\sqrt{n}-$consistent for β_0^+ and asymptotically normal, both under a regular and a singular design
- $\hat{\beta}_n^+$ is not consistent under a nearly-singular design, because $\boldsymbol{Q}_n$ is not a rank-consistent estimator of $\boldsymbol{Q}_0$

Consistent modified Ridgeless estimation of β_0^+

Definition (Modified Ridgeless estimator)

Modified Ridgeless estimator of parameter β_0^+ is:

$$\check{\beta}_n^+ := \arg\min_{\boldsymbol{\beta}} \{\|\beta\|_2 : \check{\boldsymbol{Q}}_n \beta = \boldsymbol{X}' \boldsymbol{Y}/n\} = \check{\boldsymbol{Q}}_n^+ \boldsymbol{X}' \boldsymbol{Y}/n \ ,$$

where $\check{\boldsymbol{Q}}_n$ is a consistent estimator of $\boldsymbol{Q}_0$ that is rank consistent:

$$\mathbb{P}(\text{Range}(\check{\boldsymbol{Q}}_n) = \text{Range}(\boldsymbol{Q}_0)) \to 1, \quad \text{as } n \to \infty$$

- $\check{\boldsymbol{Q}}_n$ obtained by truncating eigenvalues of $\boldsymbol{Q}_n$ with hard threshold $\mu_n \propto n^{-\alpha}$ for suitable $\alpha > 0$
- $\check{\beta}_n^+$ is clearly a consistent estimator of β_0^+ under both a regular and an irregular design

Asymptotic distribution of modified Ridgeless estimator

- MR1 $\boldsymbol{X}'\boldsymbol{\epsilon}/\sqrt{n} \to_d \boldsymbol{Z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}_0)$, with $\boldsymbol{\Omega}_0$ symm. pos. semi definite
- MR2 $\sqrt{n}(\boldsymbol{Q}_n - \boldsymbol{Q}_{0n}) \to_d \boldsymbol{\Theta}$, for some random matrix $\boldsymbol{\Theta}$
- MR3 $\boldsymbol{P}_0(\boldsymbol{Q}_n - \boldsymbol{Q}_{0n})\boldsymbol{P}_0^{\perp} = o_p(1/\sqrt{n})$

Proposition (Asymptotic distribution of modified Ridgeless estimator)

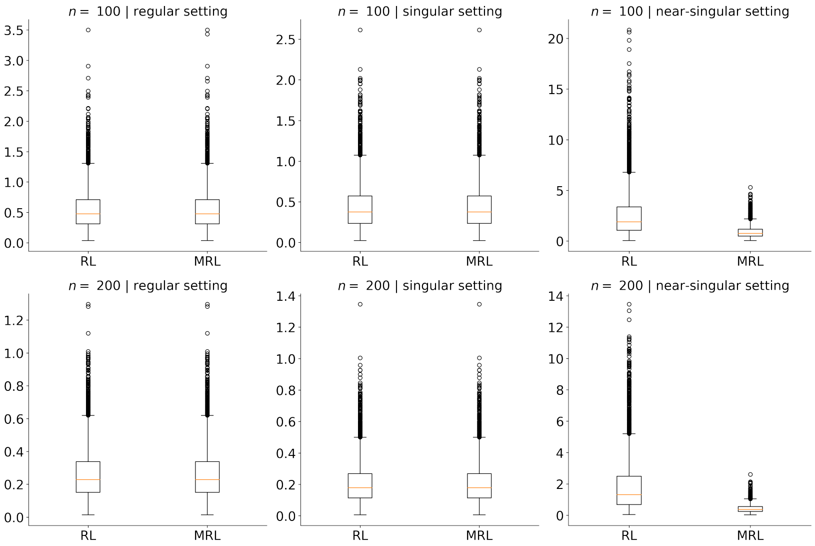
If $\boldsymbol{Q}_{0n} - \boldsymbol{Q}_0 = \frac{\boldsymbol{\Delta}}{\tau_n} + o(1/\tau_n)$, where $\tau_n \to \infty$ as $n \to \infty$, and Assumptions MR1-MR2 hold, then:

$$\sqrt{n}(\check{\boldsymbol{\beta}}_n^+ - \boldsymbol{\beta}_0^+) \to_d \boldsymbol{\eta} := \boldsymbol{P}_0(\boldsymbol{\Theta} + c\boldsymbol{\Delta})\boldsymbol{Q}_0^+\boldsymbol{\beta}_0^+ + \boldsymbol{Q}_0^+\boldsymbol{Z} \ ,$$

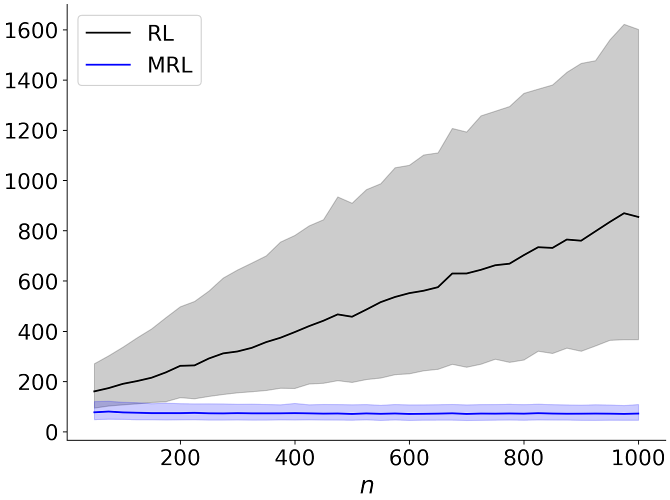
where $\sqrt{n}/\tau_n \to c \in \{0, 1\}$. Thus, if Assumption MR3 also holds and $c = 0$, then $\boldsymbol{\eta} = \boldsymbol{Q}_0^+\boldsymbol{Z}$.

- $\check{\boldsymbol{\beta}}_n^+$ is always asymptotically normally distributed whenever $vec(\boldsymbol{\Theta})$ is Gaussian

Squared errors $\left\| \hat{\beta}_n^s - \beta_0^+ \right\|_2^2$ of Ridgeless/modified Rideless



Nearly singular: normalized squared errors $n \left\| \hat{\beta}_n^s - \beta_0^+ \right\|_2^2$ of Ridgeless/modified Rideless



PEs for irregular designs

- A weighting matrix satisfying Assumption A1 under Assumptions MR1-MR2:

$$\boldsymbol{W}_n = \overline{\check{\boldsymbol{Q}}}_n := \check{\boldsymbol{Q}}_n + \boldsymbol{I} - \check{\boldsymbol{Q}}_n \check{\boldsymbol{Q}}_n^{+} \to_{\text{Pr}} \overline{\boldsymbol{Q}}_0 := \boldsymbol{Q}_0 + \boldsymbol{I} - \boldsymbol{Q}_0 \boldsymbol{Q}_0^{+}$$

Proposition (Asymptotic distribution of PEs for irregular designs)

Consider following PE:

$$prox_{\lambda_n f_n}^{\overline{\check{\boldsymbol{Q}}}_n}(\check{\boldsymbol{\beta}}_n^{+}) = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ \frac{1}{2} \left\| \check{\boldsymbol{\beta}}_n^{+} - \boldsymbol{\beta} \right\|_{\overline{\check{\boldsymbol{Q}}}_n}^{2} + \lambda_n f_n(\boldsymbol{\beta}) \right\}$$

If Assumptions MR1, MR2 and A3 hold, then:

$$\sqrt{n} \left(prox_{\lambda_n f_n}^{\overline{\check{\boldsymbol{Q}}}_n}(\check{\boldsymbol{\beta}}_n^{+}) - \boldsymbol{\beta}_0^{+} \right) \to_d prox_{g_0'(\boldsymbol{\beta}_0^{+};\cdot)}^{\overline{\boldsymbol{Q}}_0}(\boldsymbol{\eta}) = \left(Id - P_{\partial g_0(\boldsymbol{\beta}_0^{+})}^{\overline{\boldsymbol{Q}}_0} \right)(\boldsymbol{\eta})$$

Oracle PEs for irregular designs

Proposition (Oracle PE for irregular designs)

Given Adaptive Lasso penalty $f_n(\boldsymbol{\beta}) = \sum_{i=1}^{p} |\beta_i| / |\check{\beta}_{in}^{+}|$, *define PE:*

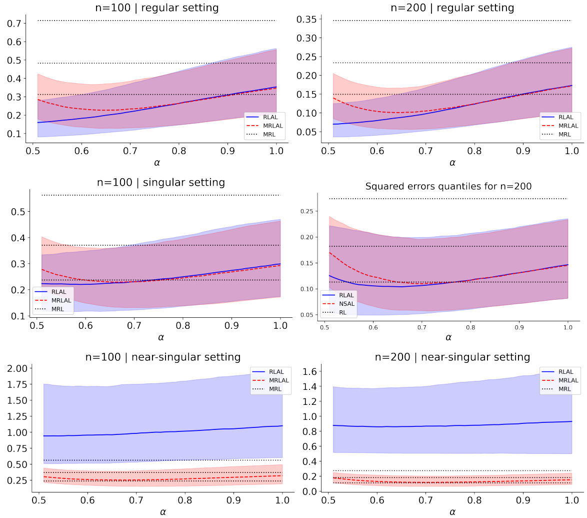
$$\operatorname{prox}_{\lambda_n f_n}^{\overline{\check{\boldsymbol{Q}}}_n}(\check{\boldsymbol{\beta}}_n^{+}) = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ \frac{1}{2} \left\| \check{\boldsymbol{\beta}}_n^{+} - \boldsymbol{\beta} \right\|_{\check{\boldsymbol{Q}}_n}^{2} + \lambda_n f_n(\boldsymbol{\beta}) \right\}$$

Let Assumptions A3 and MR1-MR3 hold with $c = 0$. *If* $\lambda_n \sqrt{n} \to 0$ *and* $\lambda_n n \to +\infty$ *then:*

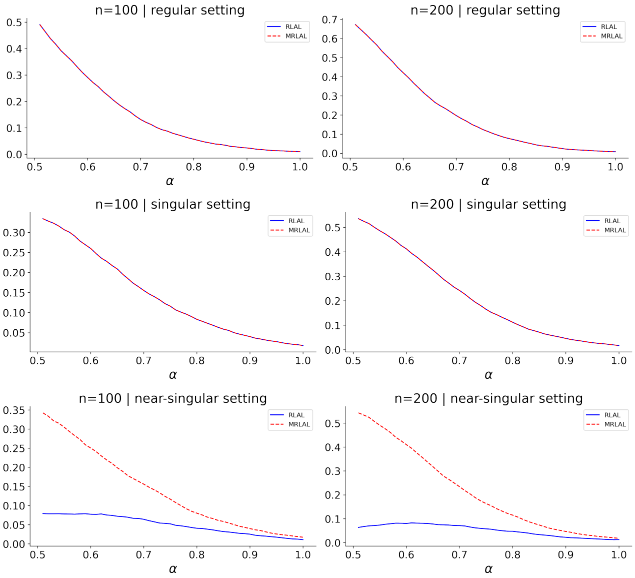
$$\sqrt{n} \begin{pmatrix} (\operatorname{prox}_{\lambda_n f_n}^{\overline{\check{\boldsymbol{Q}}}_n}(\check{\boldsymbol{\beta}}_n^{+}) - \boldsymbol{\beta}_0^{+})_{\mathcal{A}} \\ (\operatorname{prox}_{\lambda_n f_n}^{\overline{\check{\boldsymbol{Q}}}_n}(\check{\boldsymbol{\beta}}_n^{+}) - \boldsymbol{\beta}_0^{+})_{\mathcal{A}^c} \end{pmatrix} \to_d \begin{pmatrix} N\left((\mathbf{0})_{\mathcal{A}}, \sigma_0^2[(\boldsymbol{Q}_0)_{\mathcal{A}}]^{+}\right) \\ (\mathbf{0})_{\mathcal{A}_c} \end{pmatrix}.$$

Moreover, the consistent variable selection property holds.

Squared errors $\left\|\hat{\beta}_n^s - \beta_0^+\right\|_2^2$ for RLAL and MRLAL vs. $\lambda_n = n^{-\alpha}$ $[\alpha \in (0.5, 1)]$



Detection probabilities $\mathbb{P}(\hat{\mathcal{A}}_n = \mathcal{A})$ of RLAL and MRLAL vs. $\lambda_n = n^{-\alpha}$ $[\alpha \in (0.5, 1)]$



Conclusions

- A convenient class of PEs built with minimum distance corrections defined through smooth proximal operators
- A unifying convex analysis framework characterizing PEs' asymptotic properties:
  - Asymptotic distribution, Oracle property,...
- Oracle PE of minimum norm parameter in linear regression models with an irregular design
- Extensions:
  - Instrumental variables proximal estimation and inference under weak instruments
  - Estimation of stochastic discount factors in economies with nearly redundant payoffs

Thank you!