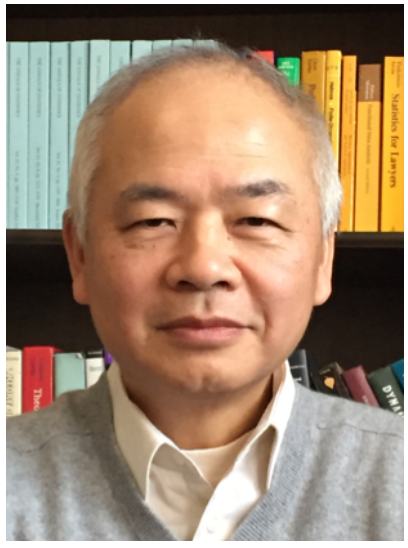
Degrees-of-freedom, asymptotic normality and risk estimation for high-dimensional regularized estimators

Pierre C Bellec, Rutgers University

Nov 11, 2020



Joint work with Cun-Hui Zhang (Rutgers).

- *Second order Poincaré inequalities and de-biasing arbitrary convex regularizers*. P.B. and Cun-Hui Zhang arXiv:1912.11943
- *Out-of-sample error estimate for robust M-estimators with convex penalty*. P.B. arXiv:2008.11840

High-dimensional statistics

- n data points $(\boldsymbol{x}_i, Y_i,\ i = 1, ..., n)$
- p covariates, $\boldsymbol{x}_i \in \mathbb{R}^p$

$$p \geq n, \qquad p \geq cn \qquad p \geq n^{\alpha}$$

In this talk:

- Linear model $Y_i = \boldsymbol{x}_i^\top \beta + \epsilon_i$ for unknown β
- $p/n \leq \gamma$ for some constant γ (or $p/n \to \gamma'$)

M-estimators and regularization

$$\hat{\beta} = \underset{\boldsymbol{b} \in \mathbb{R}^p}{\arg\min} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(\boldsymbol{x}_i^\top \boldsymbol{b}, Y_i) + \text{regularizer}(\boldsymbol{b}) \right\}$$

for some loss $\ell(\cdot, \cdot)$ and regularization penalty.

Typically in the linear model with Gaussian noise and without contamination,

$$\hat{\beta} = \underset{\boldsymbol{b} \in \mathbb{R}^p}{\arg\min} \left\{ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|^2/(2n) + g(\boldsymbol{b}) \right\}$$

with the least-squares loss for some convex penalty g convex.

Example

- Lasso, Elastic-Net
- Bridge $g(\boldsymbol{b}) = \sum_{j=1}^{p} |b_j|^c$
- Group-Lasso
- Nuclear Norm penalty
- Sorted L1 penalty (SLOPE)

Different goals, different scales

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{b}\in\mathbb{R}^p} \left\{ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|^2/(2n) + g(\boldsymbol{b}) \right\}, \qquad g \text{ convex}$$

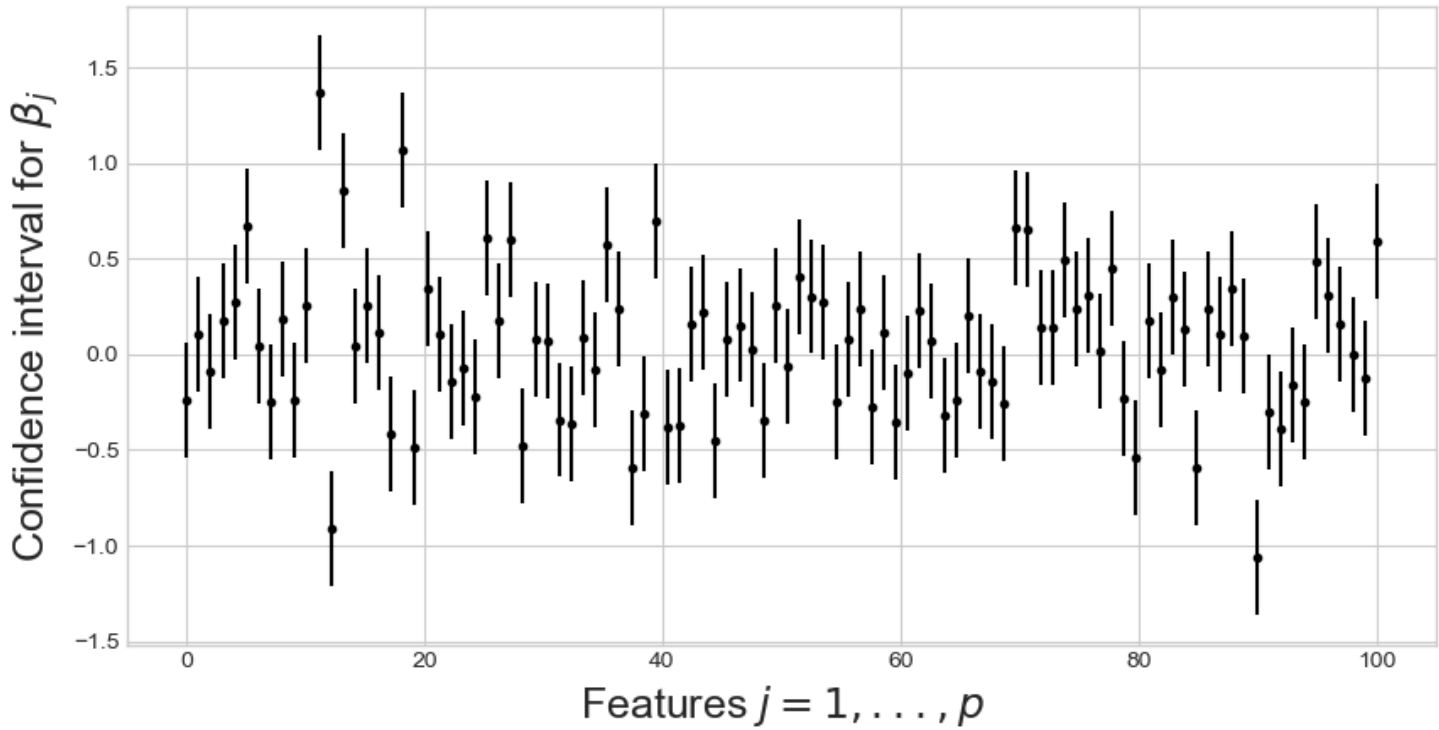
1. Design of regularizer g with intuition about complexity, structure
   - convex relaxation of unknown structure (sparsity, low-rank)
   - ℓ_1 balls are spiky at sparse vectors
2. Upper and lower bounds on the risk of $\hat{\boldsymbol{\beta}}$:
$$c r_n \le \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2 \le C r_n.$$
3. Characterization of the risk
$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2 = r_n(1 + o_P(1))$$
under some asymptotics, e.g., $p/n \to \gamma$ or $s\log(p/s)/n \to 0$.
4. Asymp. distribution in fixed direction $\boldsymbol{a}_0 \in \mathbb{R}^p$ (resp $\boldsymbol{a}_0 = \boldsymbol{e}_j$) and confidence interval for $\boldsymbol{a}_0^\top \boldsymbol{\beta}$ (resp β_j)

$$\sqrt{n}\boldsymbol{a}_0^\top(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \to^? N(0, V_0), \qquad \sqrt{n}(\widehat{\beta}_j - \beta_j) \to^? N(0, V_j).$$

Focus of today I: Confidence interval in the linear model

based on convex regularized estimators of the form

$$\hat{\beta} = \arg\min_{\boldsymbol{b} \in \mathbb{R}^p} \left\{ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|^2/(2n) + g(\boldsymbol{b}) \right\}, \qquad g \text{ convex}$$



$\sqrt{n}(\hat{b}_j - \beta_j) \Rightarrow N(0, V_j)$, $\qquad \beta_j$ unknown parameter of interest

Focus of today II: Estimation of the out-of-sample error and of σ^2

based on convex regularized estimators of the form

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{b}\in\mathbb{R}^p} \left\{ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|^2/(2n) + g(\boldsymbol{b}) \right\}, \qquad g \text{ convex}$$

Design $\boldsymbol{X}$ with iid $N(0, \boldsymbol{\Sigma})$ rows, known $\boldsymbol{\Sigma}$, noise $\varepsilon \sim N(0, \sigma^2 \boldsymbol{I}_n)$,

- estimation of noise variance σ^2
- estimation of out-of-sample error
  $\|\boldsymbol{\Sigma}^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|^2 = \mathbb{E}[((\widehat{\beta} - \beta)^\top \boldsymbol{x}_{new})^2 | (Y_1, \boldsymbol{x}_1), ..., (Y_n, \boldsymbol{x}_n)]$

(Focus I) Confidence interval in the linear model

Design $\boldsymbol{X}$ with iid $N(0, \boldsymbol{\Sigma})$ rows, known $\boldsymbol{\Sigma}$, noise $\varepsilon \sim N(0, \sigma^2 \boldsymbol{I}_n)$,

$$\boldsymbol{y} = \boldsymbol{X}\beta + \varepsilon, \qquad \text{and a given initial estimator } \hat{\beta}.$$

Goal: Inference for $\theta = \boldsymbol{a}_0^\top \beta$, projection in direction $\boldsymbol{a}_0$

Examples:

- $\boldsymbol{a}_0 = \boldsymbol{e}_j$, interested in inference on the j-th coefficient β_j
- $\boldsymbol{a}_0 = \boldsymbol{x}_{new}$ where $\boldsymbol{x}_{new}$ is the characteristics of a new patient, inference for $\boldsymbol{x}_{new}{}^\top \beta$.

De-biasing, confidence intervals for the Lasso

Confidence intervals for low dimensional parameters in high dimensional linear models

CH **Zhang**, SS **Zhang** - Journal of the Royal Statistical Society …, 2014 - Wiley Online Library

The purpose of this paper is to propose methodologies for statistical inference of low dimensional parameters with high dimensional data. We focus on constructing confidence intervals for individual coefficients and linear combinations of several of them in a linear …

☆ 99 Cited by 591 Related articles All 17 versions

On asymptotically optimal confidence regions and tests for high-dimensional models

…, P Bühlmann, Y **Ritov**, R **Dezeure** - The Annals of …, 2014 - projecteuclid.org

We propose a general method for constructing confidence intervals and statistical tests for single or low-dimensional components of a large parameter vector in a high-dimensional model. It can be easily adjusted for multiplicity taking dependence among tests into account …

☆ 99 Cited by 668 Related articles All 17 versions

[PDF] **Confidence intervals** and hypothesis testing for high-dimensional regression

A **Javanmard**, A **Montanari** - The Journal of Machine Learning Research, 2014 - jmlr.org

Fitting high-dimensional statistical models often requires the use of non-linear parameter estimation procedures. As a consequence, it is generally impossible to obtain an exact characterization of the probability distribution of the parameter estimates. This in turn implies that it is extremely challenging to quantify the uncertainty associated with a certain parameter estimate. Concretely, no commonly accepted procedure exists for computing classical measures of uncertainty and statistical significance as confidence intervals or …

☆ 99 Cited by 501 Related articles All 13 versions »

(Focus I) Confidence interval in the linear model

Design $\boldsymbol{X}$ with iid $N(0, \boldsymbol{\Sigma})$ rows, known $\boldsymbol{\Sigma}$, noise $\varepsilon \sim N(0, \sigma^2 \boldsymbol{I}_n)$,

$$\boldsymbol{y} = \boldsymbol{X}\beta + \varepsilon, \qquad \text{and a given initial estimator } \hat{\beta}.$$

Goal: Inference for $\theta = \boldsymbol{a}_0^\top \beta$, projection in direction $\boldsymbol{a}_0$

Examples:

- $\boldsymbol{a}_0 = \boldsymbol{e}_j$, interested in inference on the j-th coefficient β_j
- $\boldsymbol{a}_0 = \boldsymbol{x}_{new}$ where $\boldsymbol{x}_{new}$ is the characteristics of a new patient, inference for ${\boldsymbol{x}_{new}}^\top \beta$.

De-biasing: construct an unbiased estimate in the direction $\boldsymbol{a}_0$

i.e., find a correction such that $[\boldsymbol{a}_0^\top \hat{\beta} - \text{correction}]$ is an unbiased estimator of $\boldsymbol{a}_0^\top \beta^*$

Existing results

Lasso

- Zhang and Zhang (2014) ($s \log(p/s)/n \to 0$)
- Javanmard and Montanari (2014a) ; Javanmard and Montanari (2014b) ; Javanmard and Montanari (2018) ($s \log(p/s)/n \to 0$)
- Van de Geer et al. (2014) ($s \log(p/s)/n \to 0$)
- Bayati and Montanari (2012) ; Miolane and Montanari (2018) ($p/n \to \gamma$)

Beyond Lasso?

- Robust M-estimators El Karoui et al. (2013) Lei, Bickel, and El Karoui (2018) Donoho and Montanari (2016) ($p/n \to \gamma$)
- Celentano and Montanari (2019) symmetric convex penalty and ($\boldsymbol{\Sigma} = \boldsymbol{I}_p$, $p/n \to \gamma$), using Approximate Message Passing ideas from statistical physics
- logistic regression Sur and Candès (2018) ($\boldsymbol{\Sigma} = \boldsymbol{I}_p$, $p/n \to \gamma$)

Degrees-of-freedom of estimator

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{b}\in\mathbb{R}^p}{\arg\min} \left\{ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|^2/(2n) + g(\boldsymbol{b}) \right\}$$

- then $\boldsymbol{y} \to \boldsymbol{X}\hat{\boldsymbol{\beta}}$ for fixed $\boldsymbol{X}$ is 1-Lipscthiz
- the Jacobian of $\boldsymbol{y} \mapsto \boldsymbol{X}\hat{\boldsymbol{\beta}}$ exists everywhere (Rademacher's theorem)

$$\hat{\mathrm{df}} = \mathrm{trace}\, \nabla(\boldsymbol{y} \mapsto \boldsymbol{X}\hat{\boldsymbol{\beta}}), \qquad \hat{\mathrm{df}} = \mathrm{trace}\left[\boldsymbol{X}\frac{\partial \hat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{y})}{\partial \boldsymbol{y}}\right].$$

used for instance in Stein's Unbiased Risk Estimate (SURE).

The Jacobian matrix $\hat{\boldsymbol{H}}$ is also useful. $\hat{\boldsymbol{H}}$ is always symmetric[1]

$$\hat{\boldsymbol{H}} = \boldsymbol{X}\frac{\partial \hat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{y})}{\partial \boldsymbol{y}} \qquad \in \mathbb{R}^{n\times n}$$

[1] P.C.B and C.-H. Zhang (2019) *Second order Poincaré inequalities and de-biasing arbitrary convex regularizers when* $p/n \to \gamma$

Example for $\hat{\text{df}}$:

Lasso

$\hat{\text{df}} = |\hat{S}|$ where $\hat{S} = \{j = 1, ..., p : \hat{\beta}_j \neq 0\}$

Slope

$\hat{\text{df}} = |\{\hat{\beta}_j, j = 1, ..., p\} \setminus \{0\}|$ number of distinct nonzero values
Minami (2020)

Elastic-Net with ℓ_2 parameter μ

$\hat{\text{df}} = \text{trace}[\boldsymbol{X}_{\hat{S}}(\boldsymbol{X}_{\hat{S}}^\top \boldsymbol{X}_{\hat{S}} + n\mu \boldsymbol{I}_p)^{-1} \boldsymbol{X}_{\hat{S}}^\top]$

No closed form?

- Closed form also available for the Group-Lasso for instance.
- If no closed form solution, possibility to approximate $\hat{\text{df}}$ efficiently

Isotropic design, any g, $p/n \to \gamma$ (B. and Zhang, 2019)

Assumptions

- Sequence of linear regression problems $\boldsymbol{y} = \boldsymbol{X}\beta + \varepsilon$
- with $n, p \to +\infty$ and $p/n \to \gamma \in (0, \infty)$,
- $g : \mathbb{R}^p \to \mathbb{R}$ coercive convex penalty, strongly convex if $\gamma \geq 1$.
- Rows of $\boldsymbol{X}$ are iid $N(\boldsymbol{0}, \boldsymbol{I}_p)$ and
- Noise $\varepsilon \sim N(0, \sigma^2 \boldsymbol{I}_n)$ is independent of $\boldsymbol{X}$.

Isotropic design, any penalty g, $p/n \to \gamma$

Theorem (B. and Zhang, 2019)

$$\hat{\boldsymbol{\beta}} = \operatorname*{arg\,min}_{\boldsymbol{b} \in \mathbb{R}^p} \Big\{ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|^2/(2n) + g(\boldsymbol{b}) \Big\}$$

- $\beta_j = \langle \boldsymbol{e}_j, \boldsymbol{\beta} \rangle$ parameter of interest
- $\hat{\boldsymbol{H}} = \boldsymbol{X}(\partial/\partial \boldsymbol{y})\hat{\boldsymbol{\beta}}, \qquad \hat{\mathrm{df}} = \operatorname{trace} \hat{\boldsymbol{H}},$
- $\hat{V}(\beta_j) = \|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2 + \operatorname{trace}[(\hat{\boldsymbol{H}} - \boldsymbol{I}_n)^2](\widehat{\beta}_j - \beta_j)^2.$

Then there exists a subset $J_p \subset [p]$ of size at least $(p - \log\log p)$ s.t.

$$\sup_{j \in J_p} \Big| \mathbb{P}\Big(\frac{(n - \hat{\mathrm{df}})(\widehat{\beta}_j - \beta_j) + \boldsymbol{e}_j^\top \boldsymbol{X}^\top (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})}{\hat{V}(\beta_j)^{1/2}} \le t \Big) - \Phi(t) \Big| \to 0.$$

Resulting 0.95 confidence interval

$$\hat{CI} = \left\{ \theta \in \mathbb{R} : \left| \frac{(n - \hat{\text{df}})(\hat{\beta}_j - \beta_j) + \boldsymbol{e}_j^\top \boldsymbol{X}^\top (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})}{\hat{V}(\beta_j)} \right| \leq 1.96 \right\}$$

Variance approximation

For most coordinates, $\hat{V}(\beta_j) \approx \|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2$ and the length of the interval is

$$2 \cdot 1.96 \|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\| \Big/ (n - \hat{\text{df}}).$$

$$\hat{CI}_{approx} = \left\{ \theta \in \mathbb{R} : \left| \frac{(n - \hat{\text{df}})(\hat{\beta}_j - \beta_j) + \boldsymbol{e}_j^\top \boldsymbol{X}^\top (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})}{\|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|} \right| \leq 1.96 \right\}.$$

Correlated design, any g, $p/n \to \gamma$

Assumption

- Sequence of linear regression problems $\boldsymbol{y} = \boldsymbol{X}\beta + \varepsilon$
- with $n, p \to +\infty$ and $p/n \to \gamma \in (0, \infty)$,
- $g : \mathbb{R}^p \to \mathbb{R}$ coercive convex penalty, strongly convex if $\gamma \geq 1$.
- Rows of $\boldsymbol{X}$ are iid $N(\boldsymbol{0}, \boldsymbol{\Sigma})$ and
- Noise $\varepsilon \sim N(0, \sigma^2 \boldsymbol{I}_n)$ is independent of $\boldsymbol{X}$.

Also possible to obtain similar confidence intervals for correlated design

e.g.:

- ℓ_1-penalty and $\boldsymbol{\Sigma} \neq \boldsymbol{I}_p$
- Slope penalty and $\boldsymbol{\Sigma} \neq \boldsymbol{I}_p$

Simulations using the approximation $\hat{V}(\theta) \approx \|\boldsymbol{y} - \boldsymbol{X}\hat{\beta}\|^2$

$n = 750$, $p = 500$, correlated $\boldsymbol{\Sigma}$.

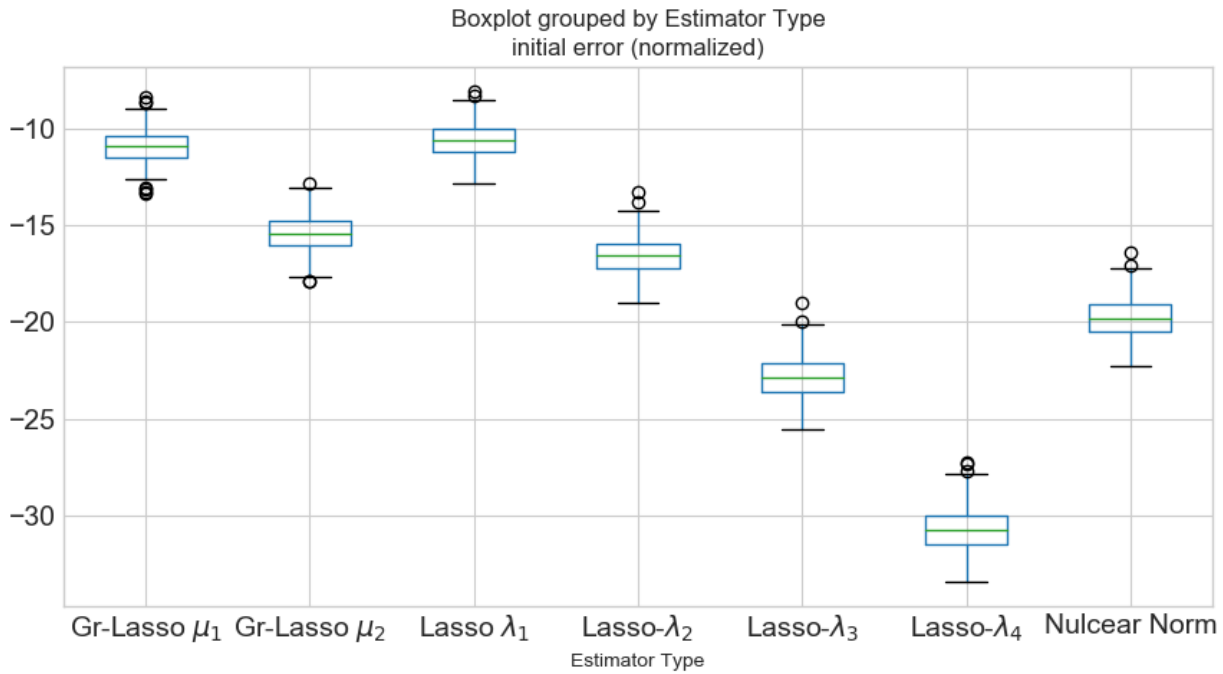
β is the vectorization of a row-sparse matrix of size 25×20.

$\boldsymbol{a}_0$ is a direction that leads to large initial bias.

Estimators: 7 different penalty functions:

- Group Lasso with tuning parameters μ_1, μ_2
- Lasso with tuning parameters $\lambda_1, ..., \lambda_4$
- Nuclear norm penalty

Boxplots of initial errors $\sqrt{n}\boldsymbol{a}_0^\top(\hat{\beta} - \beta)$ (biased!)



Simulations using the approximation $\hat{V}(\theta) \approx \|\boldsymbol{y} - \boldsymbol{X}\hat{\beta}\|^2$

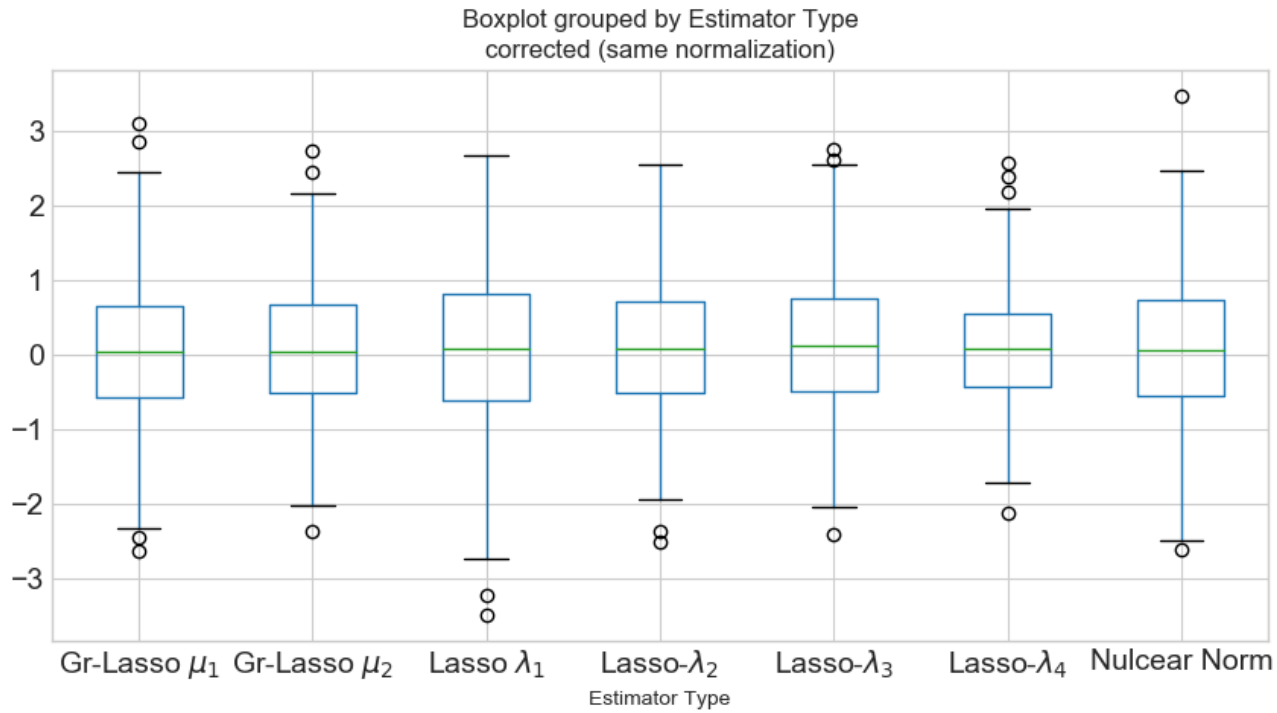
$n = 750$, $p = 500$, correlated $\boldsymbol{\Sigma}$

β is the vectorization of a row-sparse matrix of size 25×20

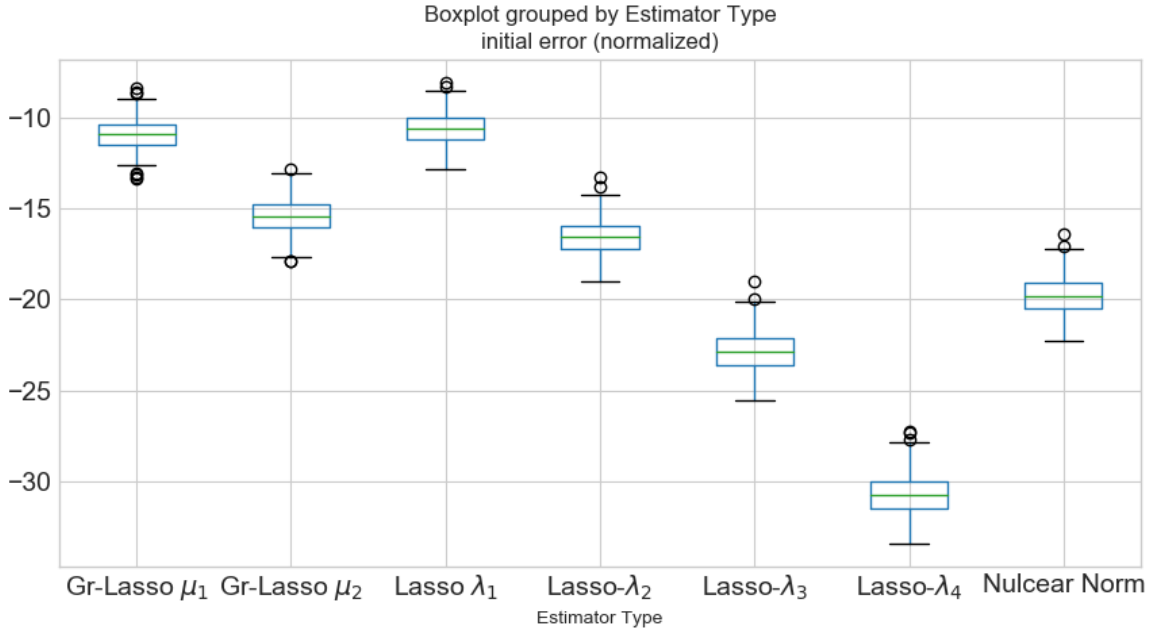
Estimators: 7 different penalty functions:

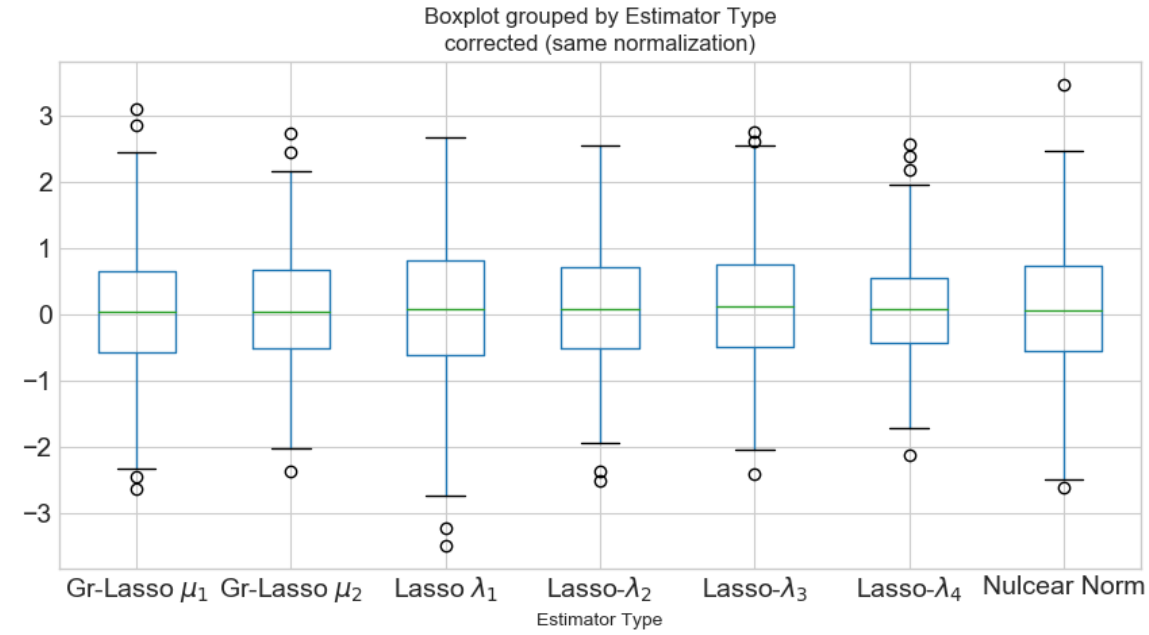
- Group Lasso with tuning parameters μ_1, μ_2
- Lasso with tuning parameters $\lambda_1, ..., \lambda_4$
- Nuclear norm penalty

Boxplots of $\sqrt{n}[\boldsymbol{a}_0^\top(\hat{\beta} - \beta) + \boldsymbol{z}_0^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\beta})]$



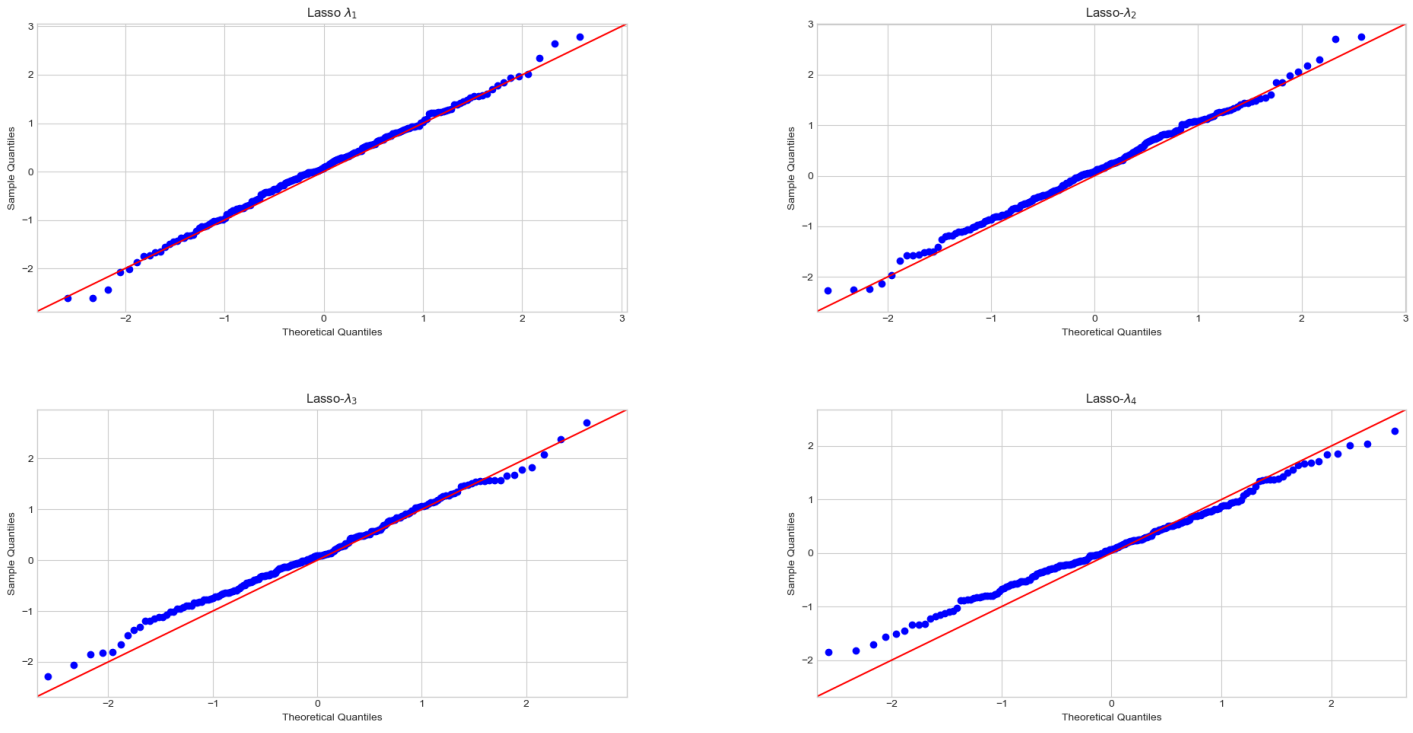
Before/after bias correction





QQ-plot, Lasso, $\lambda_1, \lambda_2, \lambda_3, \lambda_3$.

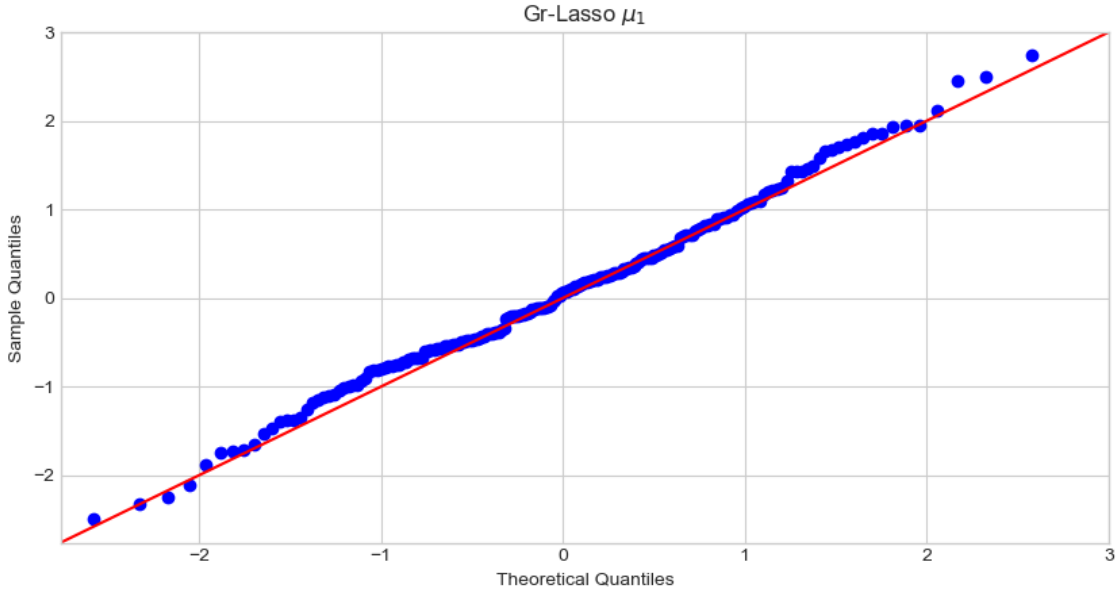
For Lasso, $\hat{\text{df}} = \left|\{j = 1, ..., p : \widehat{\beta}_j \neq 0\}\right|$.

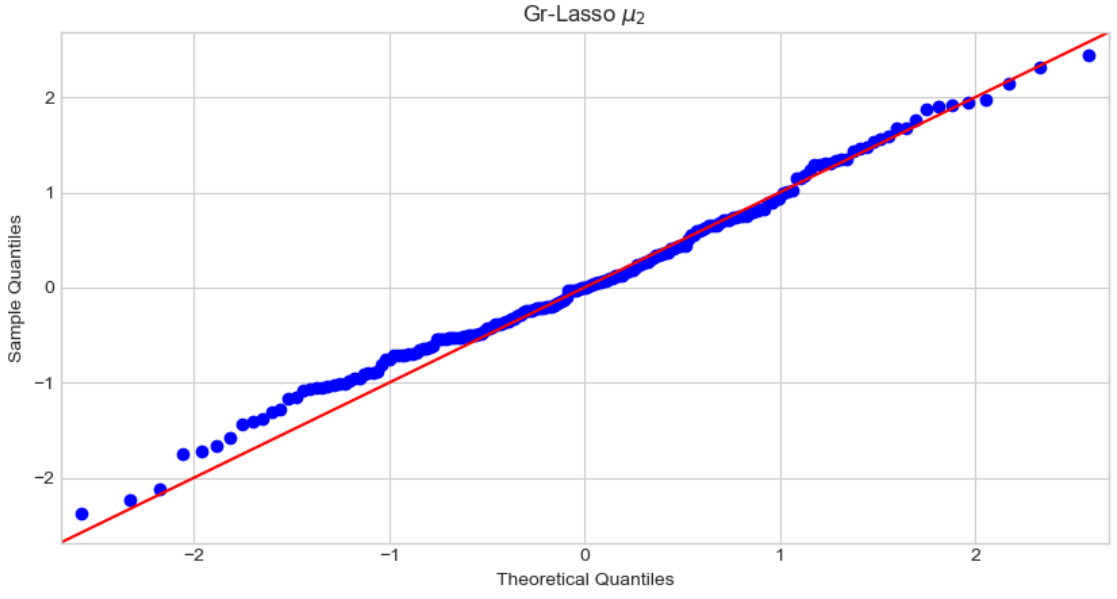


Pivotal quantity when using $\|\boldsymbol{y} - \boldsymbol{X}\hat{\beta}\|^2$ instead of $\hat{V}(\theta)$ for the variance.

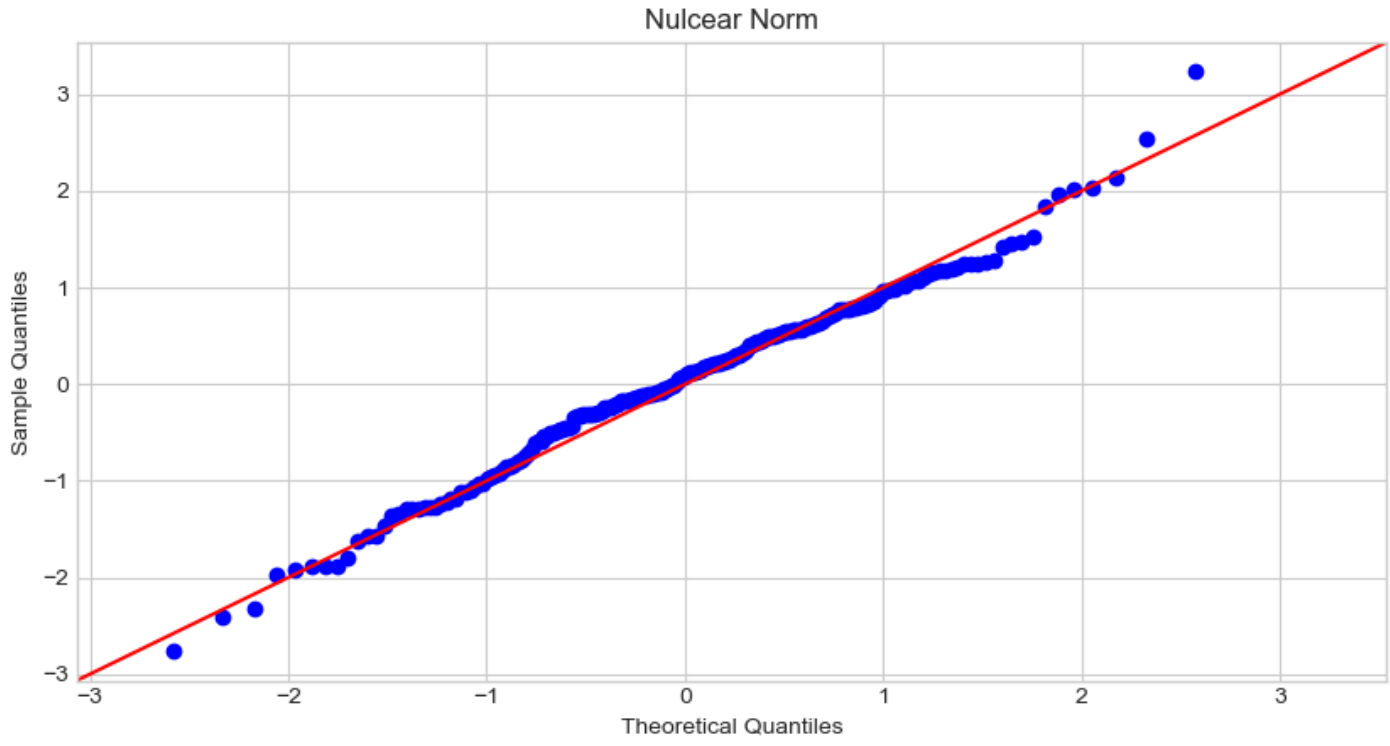
- The visible discrepancy in the last plot is fixed when using $\hat{V}(\theta)$ instead.

QQ-plot, Group Lasso, μ_1, μ_2. Explicit formula for $\hat{\text{df}}$





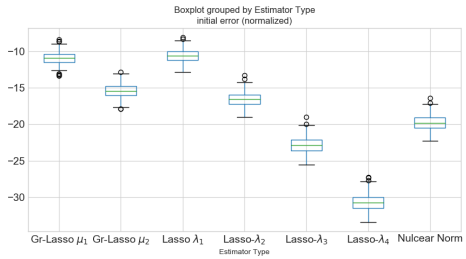
QQ-plot, Nuclear norm penalty

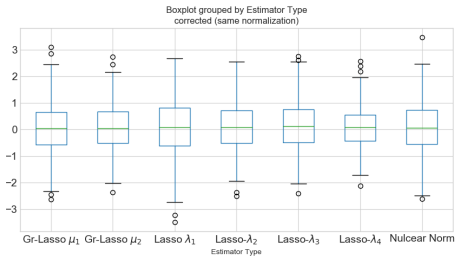


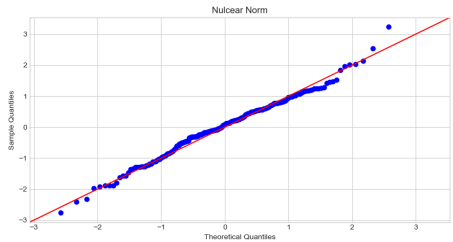
No explicit formula for $\hat{df}$ available,
although it is possible to compute numerical approximations.

(Focus I) Summary for confidence intervals:[2]

Asymptotic normality result, and valid $1-\alpha$ confidence interval by de-biasing any convex regularized M estimator.







- Asymptotics $p/n \to \gamma$
- Under Gaussian design, known covariance matrix $\boldsymbol{\Sigma}$
- Strong convexity of the penalty required if $\gamma \geq 1$; otherwise any penalty is allowed.

The effective degrees-of-freedom $\hat{\text{df}}$ plays a major role to perform the necessary de-biasing

$$\|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^{-1}\Big((n - \hat{\text{df}})(\hat{\beta}_j - \beta_j) + \boldsymbol{e}_j^\top \boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})\Big) \approx N(0,1).$$

[2] P.C.B and C.-H. Zhang (2019) *Second order Poincaré inequalities and de-biasing arbitrary convex regularizers when* $p/n \to \gamma$

Focus 2: Estimation of σ^2 and $\|\boldsymbol{\Sigma}^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|^2$

Lasso most commonly studied:
$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{b} \in \mathbb{R}^p} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|^2/(2n) + \lambda\|\boldsymbol{b}\|_1$.

With $\boldsymbol{\Sigma} = \boldsymbol{I}_p$ (isotropic design) and Bayati and Montanari (2012), Bayati, Erdogdu, and Montanari (2013), Miolane and Montanari (2018) (see also Leeb (2008) for $\lambda = 0$) provide the approximations

Generalization error
$(1 - |\hat{S}|/n)^2(\sigma^2 + \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2) \approx \|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2$

Focus 2: Estimation of σ^2 and $\|\boldsymbol{\Sigma}^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|^2$

Lasso most commonly studied:
$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{b} \in \mathbb{R}^p} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|^2/(2n) + \lambda\|\boldsymbol{b}\|_1$.

With $\boldsymbol{\Sigma} = \boldsymbol{I}_p$ (isotropic design) and Bayati and Montanari (2012), Bayati, Erdogdu, and Montanari (2013), Miolane and Montanari (2018) (see also Leeb (2008) for $\lambda = 0$) provide the approximations

Generalization error
$(1 - |\hat{S}|/n)^2(\sigma^2 + \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2) \approx \|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2$

Out-of-sample error
$(1-|\hat{S}|/n)^2\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2 \approx (\|\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2(2|\hat{S}|-p)+\|\boldsymbol{X}^\top(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})\|^2)/n^2$

Focus 2: Estimation of σ^2 and $\|\boldsymbol{\Sigma}^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|^2$

Lasso most commonly studied:
$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{b} \in \mathbb{R}^p} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|^2/(2n) + \lambda\|\boldsymbol{b}\|_1$.

With $\boldsymbol{\Sigma} = \boldsymbol{I}_p$ (isotropic design) and Bayati and Montanari (2012), Bayati, Erdogdu, and Montanari (2013), Miolane and Montanari (2018) (see also Leeb (2008) for $\lambda = 0$) provide the approximations

Generalization error
$(1 - |\hat{S}|/n)^2(\sigma^2 + \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2) \approx \|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2$

Out-of-sample error
$(1-|\hat{S}|/n)^2\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2 \approx (\|\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2(2|\hat{S}|-p)+\|\boldsymbol{X}^\top(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})\|^2)/n^2$

Noise level
$(1 - |\hat{S}|/n)^2\sigma^2 \approx (\|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2(n - 2|\hat{S}| + p) - \|\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})\|^2)/n^2$

Focus 2: Estimation of σ^2 and $\|\boldsymbol{\Sigma}^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|^2$

$\hat{\mathrm{df}}$ effective degrees-of-freedom, same as before

Theorem (P.B. 2020)

- With Gaussian design, general $\boldsymbol{\Sigma}$, either (i) strongly convex penalty and $p/n \leq \gamma \in (0, \infty)$, or (ii) $p/n \leq \gamma < 1$, or (iii) Lasso with $p/n \leq \gamma \in (0, \infty)$ and sparse $\boldsymbol{\beta}$.
- Up to an order term of order $n^{-1/2}$:

Generalization error

$$(1 - \hat{\mathrm{df}}/n)^2(\sigma^2 + \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2) \approx \|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2$$

Out-of-sample error

$$(1 - \tfrac{\hat{\mathrm{df}}}{n})^2\|\boldsymbol{\Sigma}^{\frac{1}{2}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|^2 \approx \frac{\|\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2(2\hat{\mathrm{df}}-p)+\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{X}^\top(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})\|^2}{n^2}$$

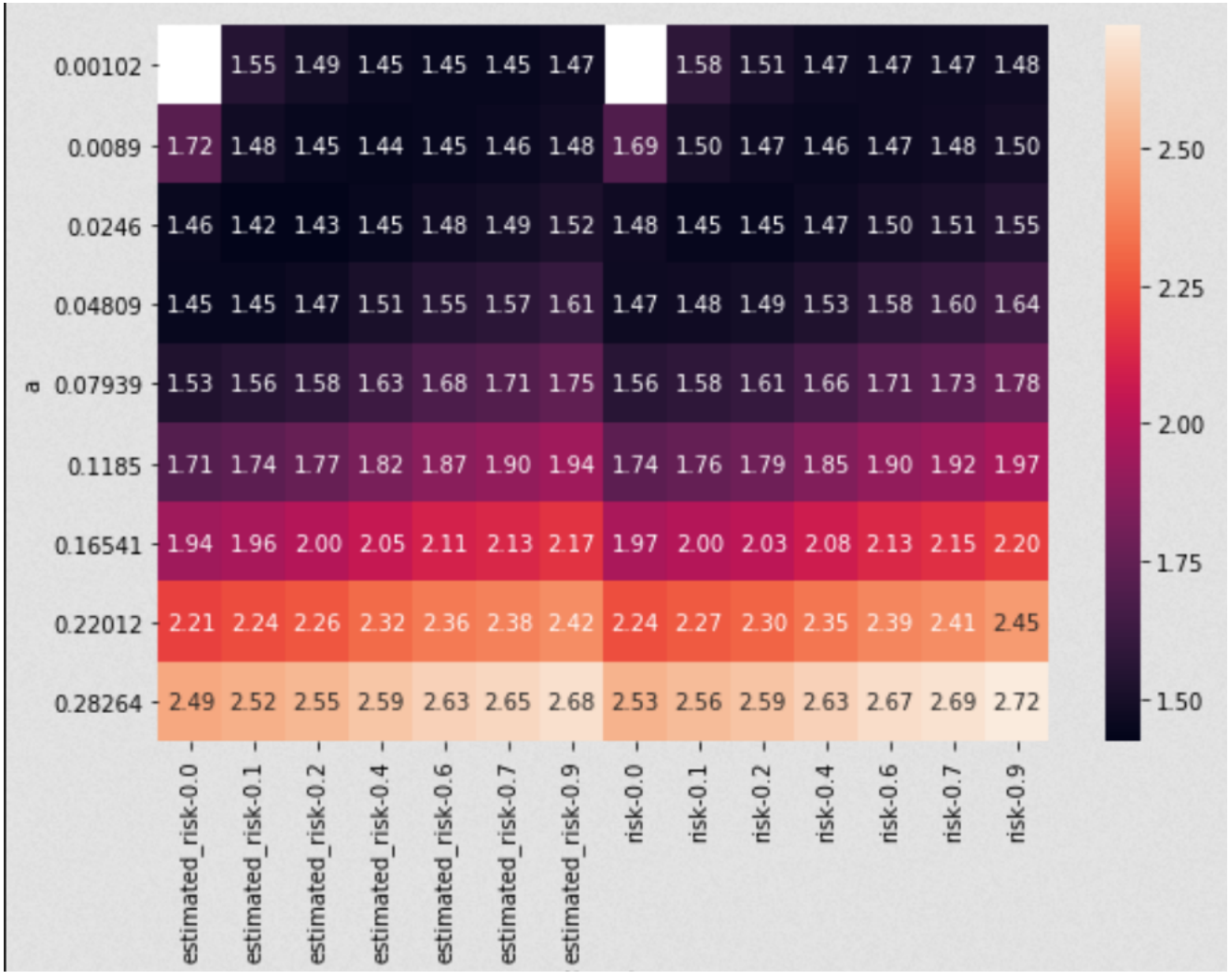
Noise level

$$(1-\hat{\mathrm{df}}/n)^2\sigma^2 \approx (\|\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2(n-2\hat{\mathrm{df}}+p)-\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}^\top(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})\|^2)/n^2$$

Parameter tuning for elastic-net

With ℓ_1 parameter λ (x-axis) ℓ_2 parameter μ (y-axis) when $\boldsymbol{\Sigma} = \boldsymbol{I}_p$:

- Left: $(1 - \hat{\mathrm{df}}/n)^{-2}\|\boldsymbol{y} - \boldsymbol{X}\hat{\beta}\|^2$ (observable)
- Right: $\|\hat{\beta} - \beta\|$ (unknown)



Extension to robust loss functions

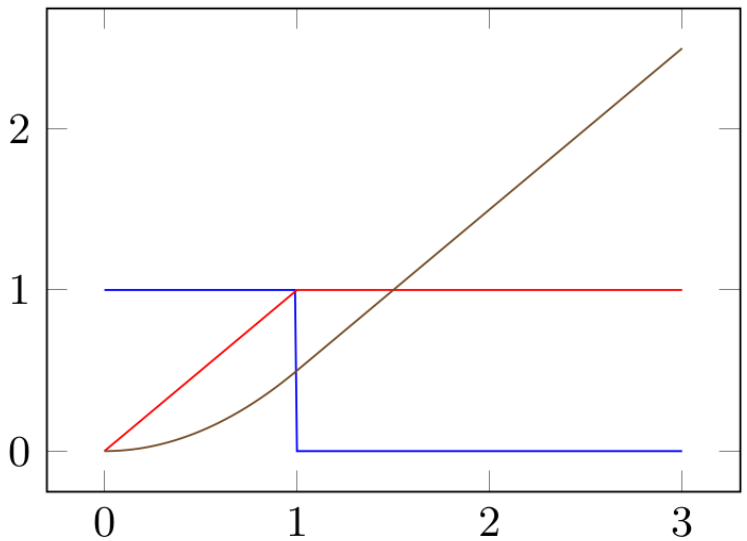
ρ robust loss function, derivative $\psi = \rho'$

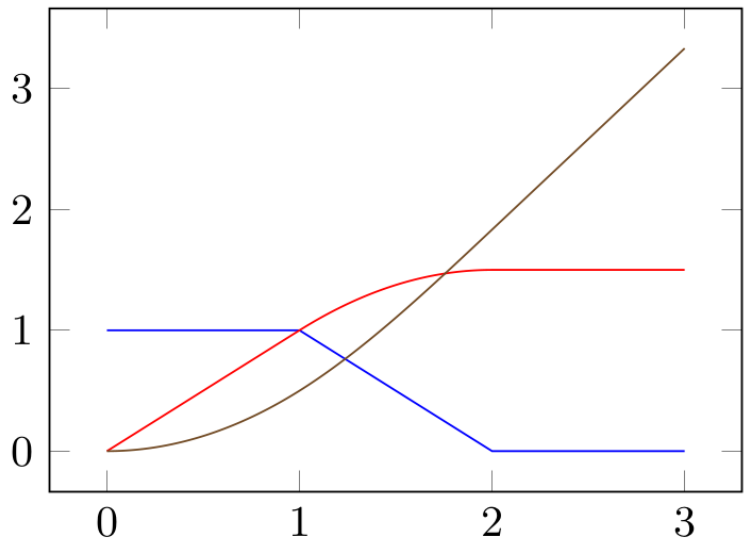
Huber loss $H(u) = \min(\frac{u^2}{2}, |u| - \frac{1}{2})$. Two parameters $\lambda, \lambda_* \geq 0$

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{b} \in \mathbb{R}^p} \sum_{i=1}^n \frac{\rho(Y_i - \boldsymbol{x}_i^\top \boldsymbol{b})}{n} + \lambda \|\boldsymbol{b}\|_1, \quad \rho(u) = \lambda_*^2 H(\lambda_*^{-1} u).$$

$u \in$	$[0,1]$	$[1,\infty)$
$\psi'_H(u)$	1	0
$\psi_H(u)$	u	1
$\rho_H(u)$	$\frac{u^2}{2}$	$u - \frac{1}{2}$

$u \in$	$[0,1]$	$[1,2]$	$[2,+\infty)$
$\psi'_0(u)$	1	$2-u$	0
$\psi_0(u)$	u	$-\frac{1}{2} + 2u - \frac{u^2}{2}$	$\frac{3}{2}$
$\rho_0(u)$	$\frac{u^2}{2}$	$\frac{1}{6} - \frac{u}{2} + u^2 - \frac{u^3}{6}$	$\frac{-7}{6} + \frac{3u}{2}$





Extension to robust loss functions: Tuning Huber Lasso

Huber loss $H(u) = \min(\frac{u^2}{2}, |u| - \frac{1}{2})$. Two parameters $\lambda, \lambda_* \geq 0$

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{b} \in \mathbb{R}^p} \sum_{i=1}^n \frac{\rho(Y_i - \boldsymbol{x}_i^\top \boldsymbol{b})}{n} + \lambda \|\boldsymbol{b}\|_1, \quad \rho(u) = \lambda_*^2 H(\lambda_*^{-1} u).$$

Inliers: $\hat{I} : \{i = 1, ..., n : \rho''(Y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}) > 0\}$

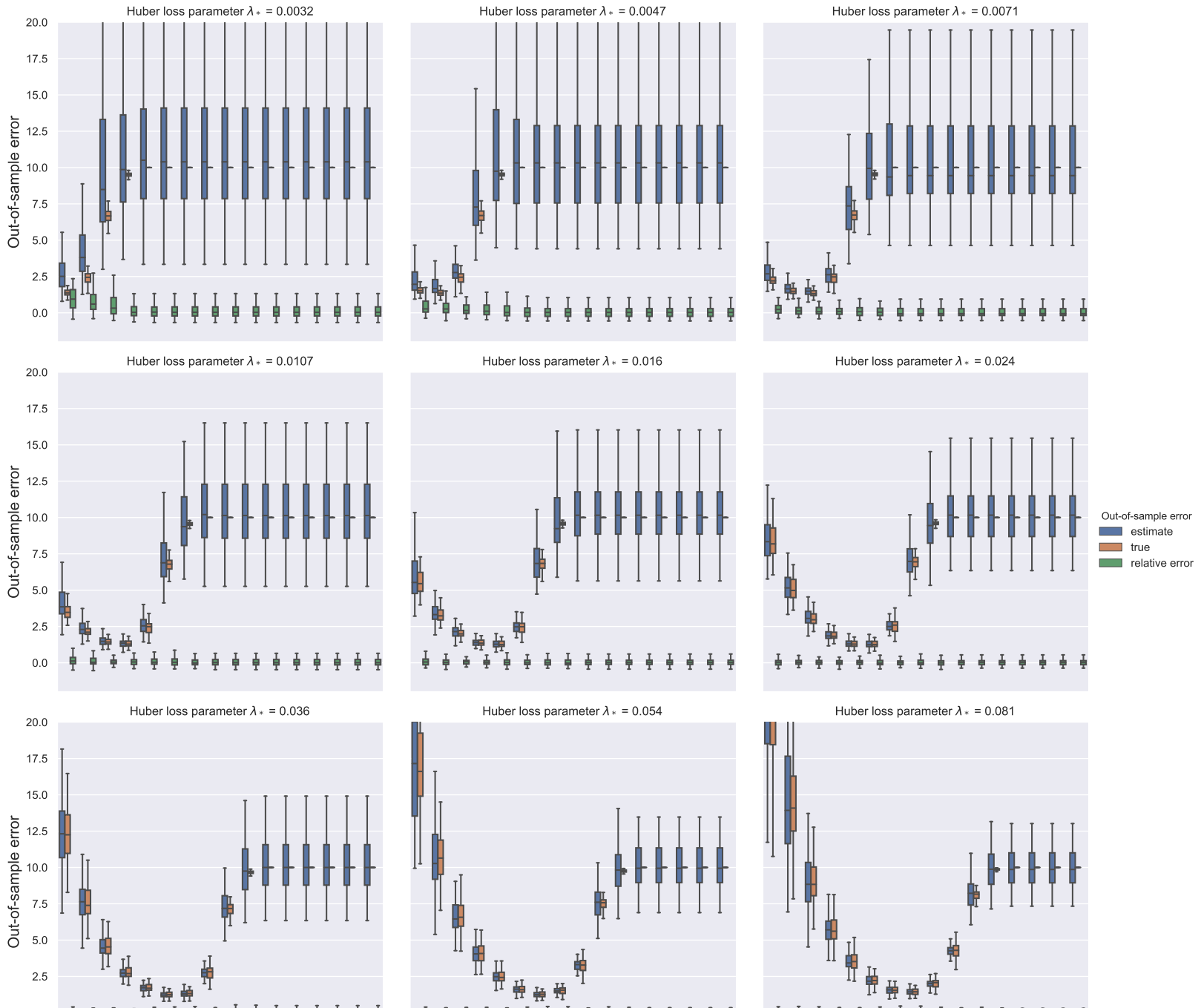
Left: Estimated Out-of-sample error

Right: True Out-of-sample error

ℓ_1-penalty parameter λ	Huber loss parameter λ_* = 0.0032	0.0047	0.0071	0.0107	0.0160	0.0240	0.0360	0.0540	0.0810	0.0032	0.0047	0.0071	0.0107	0.0160	0.0240	0.0360	0.0540	0.0810
0.0032	2.51	1.97	2.68	3.85	5.53	8.34	12.32	17.15	21.31	1.37	1.53	2.20	3.47	5.45	8.19	12.25	16.61	21.58
0.0047	3.81	1.67	1.65	2.29	3.32	5.15	7.64	10.27	13.93	2.44	1.36	1.48	2.11	3.23	4.98	7.39	10.64	14.09
0.0071	8.49	2.77	1.49	1.48	2.15	3.05	4.45	6.46	8.84	6.67	2.45	1.32	1.43	1.99	2.96	4.53	6.57	8.84
0.0107	9.86	7.28	2.63	1.34	1.39	1.85	2.73	4.05	5.71	9.51	6.70	2.47	1.28	1.35	1.83	2.70	4.08	5.61
0.016	10.51	9.75	7.36	2.54	1.28	1.30	1.72	2.48	3.43	10.00	9.52	6.73	2.48	1.26	1.29	1.69	2.43	3.52
0.024	10.39	10.32	9.95	6.88	2.47	1.25	1.22	1.62	2.17	10.00	10.00	9.54	6.79	2.48	1.26	1.25	1.59	2.21
0.036	10.39	10.32	9.35	9.37	6.84	2.50	1.30	1.25	1.58	10.00	10.00	10.00	9.56	6.84	2.60	1.32	1.27	1.55
0.054	10.39	10.32	9.45	10.20	9.24	6.98	2.76	1.50	1.44	10.00	10.00	10.00	10.00	9.58	6.95	2.82	1.56	1.44
0.081	10.39	10.32	9.45	10.14	10.16	9.45	7.18	3.31	2.01	10.00	10.00	10.00	10.00	10.00	9.62	7.17	3.28	2.07
0.1215	10.39	10.32	9.45	10.14	10.16	10.16	9.75	7.61	4.26	10.00	10.00	10.00	10.00	10.00	10.00	9.68	7.57	4.30
0.1823	10.39	10.32	9.45	10.14	10.16	10.16	10.00	9.82	8.22	10.00	10.00	10.00	10.00	10.00	10.00	10.00	9.76	8.16
0.2734	10.39	10.32	9.45	10.14	10.16	10.16	10.00	9.94	9.88	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00	9.88
0.4101	10.39	10.32	9.45	10.14	10.16	10.16	10.00	9.94	9.86	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00
0.6151	10.39	10.32	9.45	10.14	10.16	10.16	10.00	9.94	9.86	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00
0.9227	10.39	10.32	9.45	10.14	10.16	10.16	10.00	9.94	9.86	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00
1.3841	10.39	10.32	9.45	10.14	10.16	10.16	10.00	9.94	9.86	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00

$$(|\hat{I}|/n - \hat{\mathrm{df}}/n)^2 \|\boldsymbol{\Sigma}^{\frac{1}{2}}(\hat{\beta} - \beta)\|^2 \approx$$
$$(\|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2(2\hat{\mathrm{df}} - p) + \|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})\|^2)/n^2$$

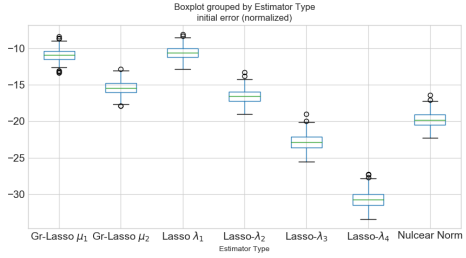
Extension to robust loss functions: Tuning Huber Lasso

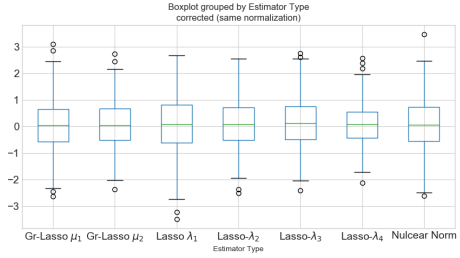


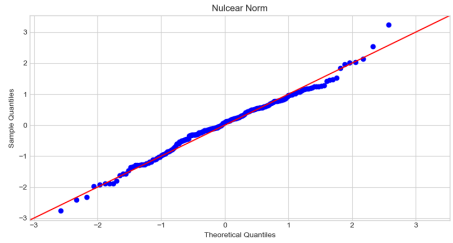
Thank you!

Focus I: Asymptotic normality result, valid $1-\alpha$ CI

$$\|\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^{-1}\Big((n-\hat{\mathrm{df}})(\hat{\beta}_j-\beta_j)+\boldsymbol{e}_j^\top\boldsymbol{X}^\top(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})\Big)\approx N(0,1).$$



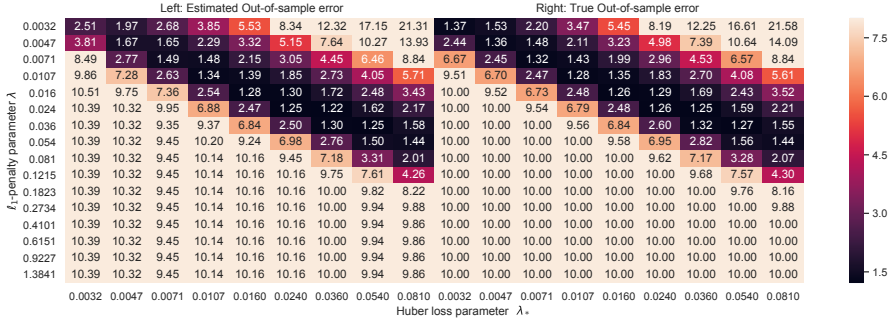




Focus 2: Estimation of σ^2 and $\|\boldsymbol{\Sigma}^{1/2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\|^2$

$$(1-\hat{\mathrm{df}}/n)^2(\sigma^2+\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|^2)\approx\|\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2$$

$$(1-\hat{\mathrm{df}}/n)^2\sigma^2\approx(\|\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2(n-2\hat{\mathrm{df}}+p)-\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}^\top(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})\|^2)/n^2$$



The effective degrees-of-freedom provides exactly the information to mitigate the high-dimensionality/undersampling

Setting

- Asymptotics $p/n \to \gamma$
- Under Gaussian design with iid rows, covariance matrix $\mathbf{\Sigma}$
- Strong convexity of the penalty required if $\gamma \geq 1$; otherwise any penalty is allowed. For the Lasso, strong convexity is not required even if $\gamma \geq 1$.

References I

Bayati, Mohsen, Murat A Erdogdu, and Andrea Montanari. 2013. "Estimating Lasso Risk and Noise Level." In *Advances in Neural Information Processing Systems*, 944–52.

Bayati, Mohsen, and Andrea Montanari. 2012. "The Lasso Risk for Gaussian Matrices." *IEEE Transactions on Information Theory* 58 (4). IEEE: 1997–2017.

Celentano, Michael, and Andrea Montanari. 2019. "Fundamental Barriers to High-Dimensional Regression with Convex Penalties." *arXiv Preprint arXiv:1903.10603*.

Donoho, David, and Andrea Montanari. 2016. "High Dimensional Robust M-Estimation: Asymptotic Variance via Approximate Message Passing." *Probability Theory and Related Fields* 166 (3-4). Springer: 935–69.

References II

El Karoui, Noureddine, Derek Bean, Peter J Bickel, Chinghway Lim, and Bin Yu. 2013. "On Robust Regression with High-Dimensional Predictors." *Proceedings of the National Academy of Sciences* 110 (36). National Acad Sciences: 14557–62.

Javanmard, Adel, and Andrea Montanari. 2014a. "Confidence Intervals and Hypothesis Testing for High-Dimensional Regression." *The Journal of Machine Learning Research* 15 (1). JMLR. org: 2869–2909.

———. 2014b. "Hypothesis Testing in High-Dimensional Regression Under the Gaussian Random Design Model: Asymptotic Theory." *IEEE Transactions on Information Theory* 60 (10). IEEE: 6522–54.

———. 2018. "Debiasing the Lasso: Optimal Sample Size for Gaussian Designs." *Ann. Statist.* 46 (6A). The Institute of Mathematical Statistics: 2593–2622. https://doi.org/10.1214/17-AOS1630.

References III

Leeb, Hannes. 2008. "Evaluation and Selection of Models for Out-of-Sample Prediction When the Sample Size Is Small Relative to the Complexity of the Data-Generating Process." *Bernoulli* 14 (3). Bernoulli Society for Mathematical Statistics; Probability: 661–90.

Lei, Lihua, Peter J Bickel, and Noureddine El Karoui. 2018. "Asymptotics for High Dimensional Regression M-Estimates: Fixed Design Results." *Probability Theory and Related Fields* 172 (3-4). Springer: 983–1079.

Minami, Kentaro. 2020. "Degrees of Freedom in Submodular Regularization: A Computational Perspective of Stein's Unbiased Risk Estimate." *Journal of Multivariate Analysis* 175. Elsevier: 104546.

Miolane, Léo, and Andrea Montanari. 2018. "The Distribution of the Lasso: Uniform Control over Sparse Balls and Adaptive Parameter Tuning." *arXiv Preprint arXiv:1811.01212*.

References IV

Sur, Pragya, and Emmanuel J Candès. 2018. "A Modern Maximum-Likelihood Theory for High-Dimensional Logistic Regression." *arXiv Preprint arXiv:1803.06964*.

Van de Geer, Sara, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. 2014. "On Asymptotically Optimal Confidence Regions and Tests for High-Dimensional Models." *The Annals of Statistics* 42 (3). Institute of Mathematical Statistics: 1166–1202.

Zhang, Cun-Hui, and Stephanie S Zhang. 2014. "Confidence Intervals for Low Dimensional Parameters in High Dimensional Linear Models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76 (1). Wiley Online Library: 217–42.