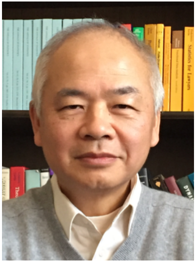
De-biasing arbitrary convex regularizers and asymptotic normality

Pierre C Bellec, Rutgers University

Mathematical Methods of Modern Statistics 2, June 2020



Joint work with Cun-Hui Zhang (Rutgers).

- *Second order Poincaré inequalities and de-biasing arbitrary convex regularizers* arXiv:1912.11943
- *De-biasing the Lasso with degrees-of-freedom adjustment.* arXiv:1902.08885.

High-dimensional statistics

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

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

For instance, linear model $Y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \epsilon_i$ for unknown $\boldsymbol{\beta}$

M-estimators and regularization

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

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

Typically in the linear model, with the least-squares loss,

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

with 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 $\hat{\boldsymbol{\beta}}$:

$$cr_n \leq \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2 \leq Cr_n.$$

3. Characterization of the risk

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2 = r_n(1 + o_P(1))$$

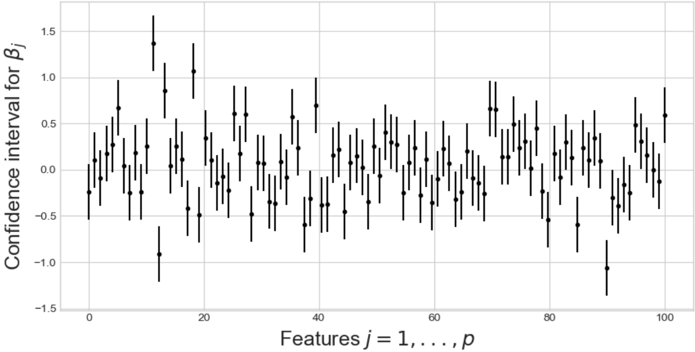
   under some asymptotics, e.g., $p/n \to \gamma$ or $s\log(p/s)/n \to 0$.
4. Asymp. distribution in fixed direction $\boldsymbol{a}_0 \in \mathbb{R}^p$ (resp $\boldsymbol{a}_0 = \boldsymbol{e}_j$) and confidence interval for $\boldsymbol{a}_0^\top\boldsymbol{\beta}$ (resp β_j)

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

Focus of today: Confidence interval in the linear model

based on convex regularized estimators of the form

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



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

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 »

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) ($p/n \to \gamma$)

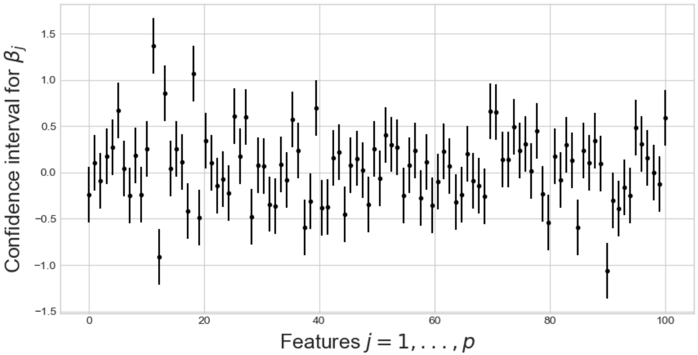
Beyond Lasso?

- 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) ($\boldsymbol{\Sigma} = \boldsymbol{I}_p$, $p/n \to \gamma$)

Focus today: General theory for confidence intervals

based on **any** convex regularized estimators of the form

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



Little or no constraint on the convex regularizer g.

Degrees-of-freedom of estimator

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

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

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

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]

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

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

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

Assumptions

- Sequence of linear regression problems $\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{\varepsilon}$
- 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 $N(\mathbf{0}, \boldsymbol{I}_p)$ and
- Noise $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \boldsymbol{I}_n)$ is independent of $\boldsymbol{X}$.

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

Theorem (B. and Zhang, 2019)

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

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

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

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

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

Assumption

- Sequence of linear regression problems $\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{\varepsilon}$
- 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 $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \boldsymbol{I}_n)$ is independent of $\boldsymbol{X}$.

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

Theorem (B. and Zhang, 2019)

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

- $\theta = \langle \boldsymbol{a}_0, \boldsymbol{\beta} \rangle$ parameter of interest
- $\hat{\boldsymbol{H}} = \boldsymbol{X}(\partial/\partial\boldsymbol{y})\hat{\beta}$, $\quad \hat{\text{df}} = \text{trace}\,\hat{\boldsymbol{H}}$,
- $\hat{V}(\theta) = \|\boldsymbol{y} - \boldsymbol{X}\hat{\beta}\|^2 + \text{trace}[(\hat{\boldsymbol{H}} - \boldsymbol{I}_n)^2](\langle \boldsymbol{a}_0, \hat{\beta}\rangle - \theta)^2$.
- Assume $\boldsymbol{a}_0^\top \boldsymbol{\Sigma} \boldsymbol{a}_0 = 1$ and set

$$\boldsymbol{z}_0 = \boldsymbol{\Sigma}^{-1}\boldsymbol{a}_0.$$

Then there exists a subset $\overline{S} \subset S^{p-1}$ with relative volume $|\overline{S}|/|S^{p-1}| \geq 1 - 2e^{-p^{0.99}}$

$$\sup_{\boldsymbol{a}_0 \in \boldsymbol{\Sigma}^{1/2}\overline{S}} \left| \mathbb{P}\Big(\frac{(n - \hat{\text{df}})(\langle \hat{\beta}, \boldsymbol{a}_0\rangle - \theta) + \langle \boldsymbol{z}_0, \boldsymbol{y} - \boldsymbol{X}\hat{\beta}\rangle}{\hat{V}(\theta)^{1/2}} \leq t \Big) - \Phi(t) \right| \to 0.$$

This applies to at least $(p - \phi_{\text{cond}}(\boldsymbol{\Sigma}) \log\log p)$ indices $j \in [p]$.

Resulting 0.95 confidence interval

$$\hat{CI} = \left\{ \theta \in \mathbb{R} : \left| \frac{(n - \hat{\mathrm{df}})(\langle \hat{\beta}, \boldsymbol{a}_0 \rangle - \theta) + \langle \boldsymbol{z}_0, \boldsymbol{y} - \boldsymbol{X}\hat{\beta} \rangle}{\hat{V}(\theta)^{1/2}} \right| \leq 1.96 \right\}$$

Variance approximation

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

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

$$\hat{CI}_{approx} = \left\{ \theta \in \mathbb{R} : \left| \frac{(n - \hat{\mathrm{df}})(\langle \hat{\beta}, \boldsymbol{a}_0 \rangle - \theta) + \langle \boldsymbol{z}_0, \boldsymbol{y} - \boldsymbol{X}\hat{\beta} \rangle}{\|\boldsymbol{y} - \boldsymbol{X}\hat{\beta}\|} \right| \leq 1.96 \right\}.$$

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

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

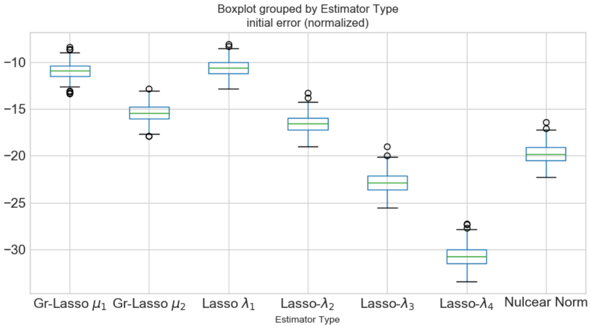
β is the vectorization of a row-sparse matrix of size 25×20.

$\boldsymbol{a}_0$ 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 $\hat{V}(\theta) \approx \|\boldsymbol{y} - \boldsymbol{X}\hat{\beta}\|^2$

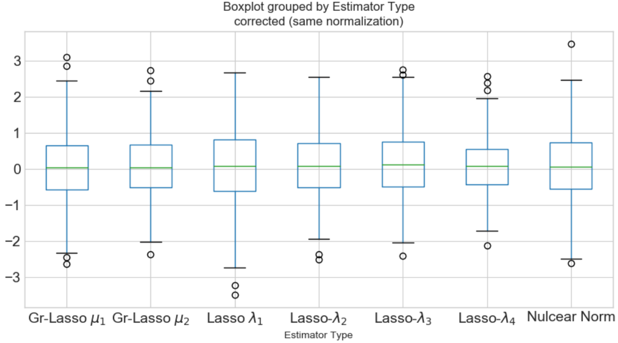
$n = 750$, $p = 500$, correlated $\boldsymbol{\Sigma}$

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

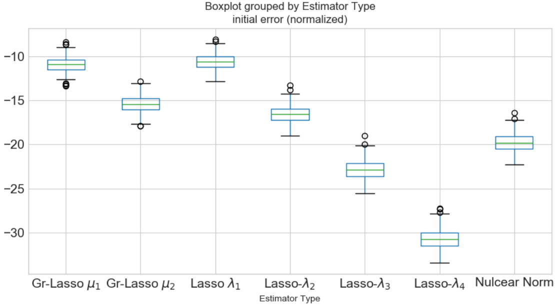
Estimators: 7 different penalty functions:

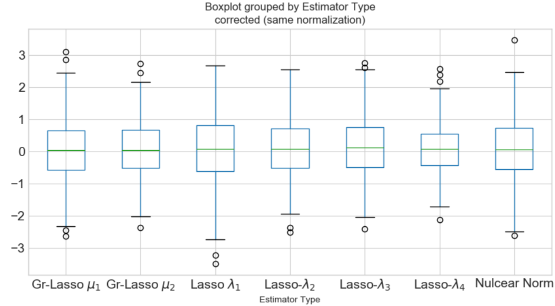
- Group Lasso with tuning parameters μ_1, μ_2
- Lasso with tuning parameters $\lambda_1, ..., \lambda_4$
- Nuclear norm penalty

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



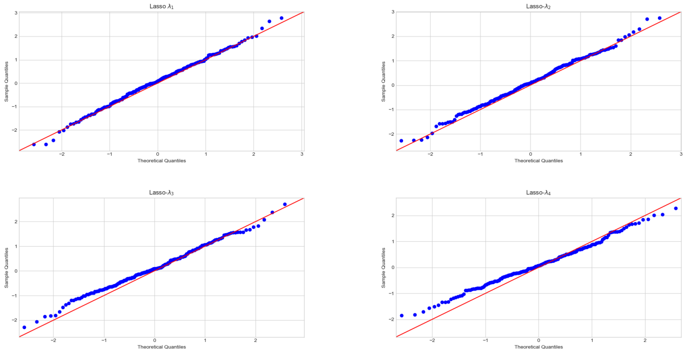
Before/after bias correction





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

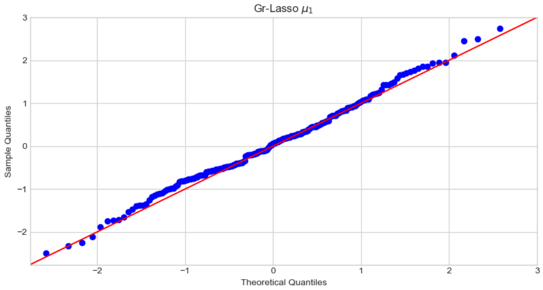
For Lasso, $\hat{\text{df}} = \left|\{j = 1, ..., p : \widehat{\beta}_j \neq 0\}\right|$.

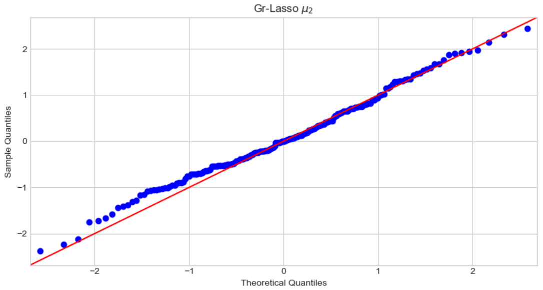


Pivotal quantity when using $\|\boldsymbol{y} - \boldsymbol{X}\hat{\beta}\|^2$ instead of $\hat{V}(\theta)$ for the variance.

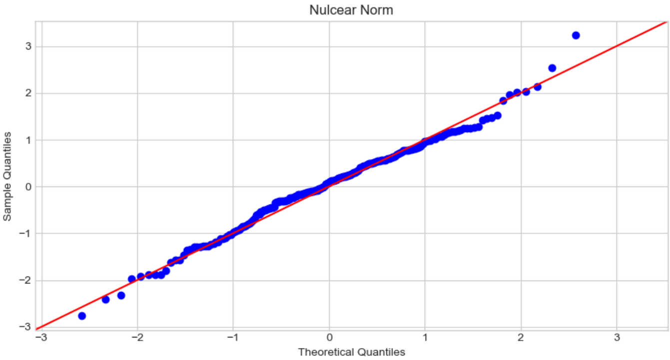
- The visible discrepancy in the last plot is fixed when using $\hat{V}(\theta)$ instead.

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





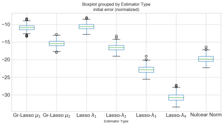
QQ-plot, Nuclear norm penalty

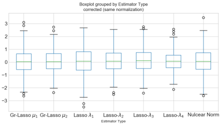


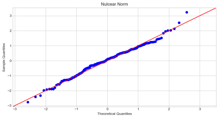
No explicit formula for $\hat{\mathrm{df}}$ available,
although it is possible to compute numerical approximations.

Summary of the main result[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 $\mathbf{\Sigma}$
- Strong convexity of the penalty required if $\gamma \geq 1$; otherwise any penalty is allowed.

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

Time-pertmitting

1. Necessity of degrees-of-freedom adjustment
2. Central Limit Theorems and Second Order Poincar'e inequalities
3. Unknown $\mathbf{\Sigma}$.

1. Necessity of degrees-of-freedom adjustment

The previous de-biasing correction features a "degrees-of-freedom" adjustment in the form of multiplication by

$$(1 - \hat{\mathrm{df}}/n)$$

or depending on the normalization, multiplication by

$$n - \hat{\mathrm{df}}.$$

Generalization, in high-dimensions, of the classical normalization by multiplying by $n - p$ to obtain unbiased estimates when $p \lll n$.

This degrees-of-freedom adjustment for the Lasso was initially motivated by statistical physics arguments[3]

[3] Javanmard and Montanari (2014b), *Hypothesis Testing in High-Dimensional Regression under the Gaussian Random Design Model: Asymptotic Theory*

Initial proposals for de-biasing the Lasso do not include the "degrees-of-freedom" adjustment

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 …

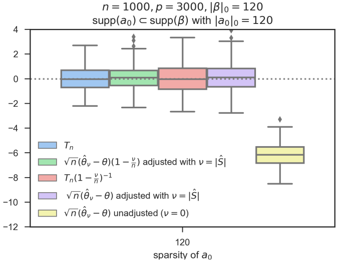
☆ 99 Cited by 501 Related articles All 13 versions »

1. Necessity of degrees-of-freedom adjustment

- Sparse linear regression $\boldsymbol{y} = \boldsymbol{X}\beta + \varepsilon$, sparsity $s_0 = \|\beta\|_0$
- $\boldsymbol{X}$ has iid $N(0, \boldsymbol{\Sigma})$ rows, noise $\varepsilon \sim N(0, \sigma^2 \boldsymbol{I}_n)$

$\hat{\theta}_\nu$ de-biasing estimate with adjustment of the form $(1 - \nu/n)$, here ν represents a possible degrees-of-freedom adjustment or absence thereof $(\nu = 0)$.

$\sqrt{n}(\hat{\theta}_\nu - \theta)$ when the initial estimator is the Lasso



The pivotal quantity for $\nu = 0$ (unadjusted) is biased. (Yellow boxplot).
The degrees-of-freedom adjustment exactly repairs this.

For $s_0 \ggg n^{2/3}$, absence of degrees-of-freedom adjustment provably leads to incorrect coverage for certain directions $\boldsymbol{a} - 0$.[4]

[4]B. and Zhang (2018): *De-Biasing The Lasso With Degrees-of-Freedom Adjustment*

2. Central Limit Theorems/Second Order Poinacr'e inequalities

If $f : \mathbb{R}^n \to \mathbb{R}^n$ and $\boldsymbol{z}_0 \sim N(0, \boldsymbol{I}_n)$, then the random variable

$$\boldsymbol{z}_0^\top f(\boldsymbol{z}_0) - \text{div} f(\boldsymbol{z}_0)$$

is close to normal when

$$\frac{\mathbb{E}\|\nabla f(\boldsymbol{z}_0)\|_F^2}{\mathbb{E}\|f(\boldsymbol{z}_0)\|^2}$$

is small[5].

- This leads to the asymptotic normal results when de-biasing convex regularizers
- Mechanically computing/bounding gradients leads to asymptotic normality results (Second Order Poincar'e inequalities, see Chatterjee (2009))

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

3. Unknown $\boldsymbol{\Sigma}$

The general theory of de-biasing/asymptotic normality for arbitrary regularizers is applicable to any penalty when $\boldsymbol{\Sigma}$ is known.

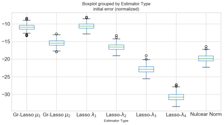
In practice, $\boldsymbol{z}_0 = \boldsymbol{\Sigma}^{-1}\boldsymbol{a}_0$ needs to be estimated.

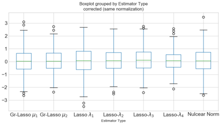
- sample splitting
- case-by-case basis for a given regularizer g
- e.g.: Nodewise Lasso. Dense and sparse $\boldsymbol{a}_0$ have to be handled differently.[6]
  - leaves open interesting problems for different regularizers

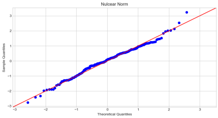
[6] B. and Zhang (2018), Section 2.2. *De-Biasing The Lasso With Degrees-of-Freedom Adjustment*

Thank you!

Asymptotic normality result, and valid $1-\alpha$ confidence interval[7] 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.

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

References I

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*.

Chatterjee, Sourav. 2009. "Fluctuations of Eigenvalues and Second Order Poincaré Inequalities." *Probability Theory and Related Fields* 143 (1-2). Springer: 1–40.

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." *The Annals of Statistics* 46 (6A). Institute of Mathematical Statistics: 2593–2622.

References III

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.

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*.

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.

References IV

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.