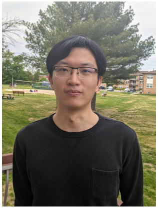
Observable adjustments for M-estimation in single index models

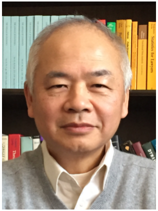
Pierre C Bellec — Rutgers University

Dec 18, ICSDS 2023, Lisbon

Collaborators



Yiwei Shen
(Former PhD student
now facebook)



Cun-Hui Zhang
(Rutgers)

Kai Tan
(Rutgers, current PhD
student)

Single index model as $n/p \to$constant

iid observations $(\boldsymbol{x}_i, y_i)_{i=1,\dots,n}$ with Gaussian feature vectors $\boldsymbol{x}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$ and response y_i

$$y_i = F(\boldsymbol{x}_i^T \boldsymbol{w}, U_i)$$

- $F : \mathbb{R}^2 \to \mathbb{R}$ is an unknown deterministic function
- $\boldsymbol{w} \in \mathbb{R}^p$ an unknown index, normalized with $\mathrm{Var}[\boldsymbol{x}_i^T \boldsymbol{w}] = \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{w}\|^2 = 1$
- U_i is a latent variable independent of $\boldsymbol{x}_i$.

Single index model as $n/p \to$constant

iid observations $(\boldsymbol{x}_i, y_i)_{i=1,\dots,n}$ with Gaussian feature vectors $\boldsymbol{x}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$ and response y_i

$$y_i = F(\boldsymbol{x}_i^T \boldsymbol{w}, U_i)$$

- $F : \mathbb{R}^2 \to \mathbb{R}$ is an unknown deterministic function
- $\boldsymbol{w} \in \mathbb{R}^p$ an unknown index, normalized with $\text{Var}[\boldsymbol{x}_i^T \boldsymbol{w}] = \|\boldsymbol{\Sigma}^{1/2} \boldsymbol{w}\|^2 = 1$
- U_i is a latent variable independent of $\boldsymbol{x}_i$.

Examples

- *Linear regression*: $F(v, u) = \|\boldsymbol{\Sigma}^{1/2} \boldsymbol{\beta}^*\| v + u$ for some $\boldsymbol{\beta}^* \in \mathbb{R}^p$, $U_i \sim N(0, \sigma^2)$ and $\boldsymbol{w} = \boldsymbol{\beta}^* / \|\boldsymbol{\Sigma}^{1/2} \boldsymbol{\beta}^*\|$.
- *Logistic regression*: $F(v, u) = 1$ if $u \leq 1/(1 + e^{-\|\boldsymbol{\beta}^*\| v})$ and 0 otherwise for some $\boldsymbol{\beta}^* \in \mathbb{R}^p$, $U_i \sim \text{Unif}[0, 1]$ and $\boldsymbol{w} = \boldsymbol{\beta}^* / \|\boldsymbol{\Sigma}^{1/2} \boldsymbol{\beta}^*\|$.
- *1-bit compressed sensing with an ϵ-proportion of bits flipped*: $F(v, u) = u\text{sign}(v)$ for $U_i \in \{-1, 1\}$ s.t. $\mathbb{P}(U_i = -1) = \varepsilon$.

Least-Squares!

$$\hat{\beta} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}$$

When $y_i | \boldsymbol{x}_i$ is nonlinear

Examples:

- Logistic model $\mathbb{E}[y_i | \boldsymbol{x}_i] = \frac{e^{x_i^T \boldsymbol{w}}}{1+e^{x_i^T \boldsymbol{w}}}$
- 1-bit compressed sensing

  $$y_i = u_i \text{sign}(\boldsymbol{x}_i^T \boldsymbol{w})$$

  with u_i random sign.
- Poisson model

Situation: Response y_i is far from linear in $x_i^T w$

Least-Squares! $\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$; $\Sigma = \boldsymbol{I}_p$ and $\|\boldsymbol{w}\| = 1$

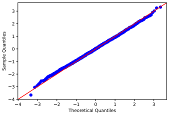
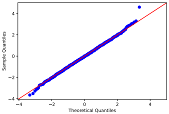
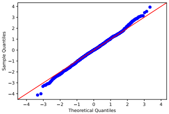
$$\hat{a}^2 = \frac{1}{n}\|\boldsymbol{X}\hat{\beta}\|^2 - \frac{p/n}{n-p}\|\boldsymbol{y} - \boldsymbol{X}\hat{\beta}\|^2 \quad \text{estimates} \quad (\boldsymbol{w}^T\hat{\beta})^2$$

QQplot $\frac{n-p}{\Omega_{jj}^{1/2}\|\boldsymbol{y}-\boldsymbol{X}\hat{\beta}\|}\left[\hat{\beta}_j - \pm\hat{a}w_j\right] \approx N(0,1) \begin{cases} \text{shrinking adjustment } \hat{a} \\ \text{variance adjustment} \end{cases}$

Least-Squares! $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$; $\boldsymbol{\Sigma} = \boldsymbol{I}_p$ and $\|\boldsymbol{w}\| = 1$

$$\hat{a}^2 = \frac{1}{n}\|\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2 - \frac{p/n}{n-p}\|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2 \quad \text{estimates} \quad (\boldsymbol{w}^T\hat{\boldsymbol{\beta}})^2$$

QQplot $\frac{n-p}{\Omega_{jj}^{1/2}\|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|}\left[\hat{\beta}_j - \pm\hat{a}w_j\right] \approx N(0,1)$ { shrinking adjustment $\hat{a}$; variance adjustment }

$\frac{p}{n} = 0.8$	Linear	Logistic $y_i \in \{0,1\}$	1-bit $y_i \in \{\pm 1\}$
$y_i \| \boldsymbol{x}_i$	$y_i \sim N(\boldsymbol{x}_i^T\boldsymbol{w}, 0.5)$	$\mathbb{E}[y_i \| \boldsymbol{x}_i] = \frac{e^{\boldsymbol{x}_i^T\boldsymbol{w}}}{1+e^{\boldsymbol{x}_i^T\boldsymbol{w}}}$	$y_i = u_i \text{sign}(\boldsymbol{x}_i^T\boldsymbol{w})$
$\hat{a}$	$.999 \pm .021$	$.407 \pm .072$	$.475 \pm .05$
$\boldsymbol{w}^T\hat{\boldsymbol{\beta}}$	$.999 \pm .027$	$-.413 \pm .033$	$.483 \pm .037$
QQplot			

For 1-Bit compressed sensing, $\mathbb{P}(u_i = -1) = 0.2 = 1 - \mathbb{P}(u_i = 1)$

M-estimator

$\hat{\beta}$ is a regularized M-estimator of the form

$$\hat{\beta}(\boldsymbol{y}, \boldsymbol{X}) = \operatorname*{argmin}_{\boldsymbol{b} \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \ell_{y_i}(\boldsymbol{x}_i^T \boldsymbol{b}) + g(\boldsymbol{b})$$

where

- $g : \mathbb{R}^p \to \mathbb{R}$ is a convex penalty function and for any $y_0 \in \mathcal{Y}$,
- the map $\ell_{y_0} : \mathbb{R} \to \mathbb{R}$, $t \mapsto \ell_{y_0}(t)$ is a convex loss function.

For a fixed y_0, the derivatives of ℓ_{y_0} are denoted by $\ell'_{y_0}(t)$ and $\ell''_{y_0}(t)$ where these derivatives exist.

- We never differentiate wrt y_0! (y_0 may be discrete)

Regime

$n/p \to \delta$ (=constant)

Ridge Logistic regression; sigmoid $\sigma(u) = 1/(1+e^{-u})$

$$\hat\beta = \operatorname*{argmin}_{\boldsymbol{b}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^n \left(\log(1+e^{\boldsymbol{x}_i^T\boldsymbol{b}}) - y_i\boldsymbol{x}_i^T\boldsymbol{b}\right) + \lambda\|\boldsymbol{b}\|^2/2$$

Define the adjustments $\hat r^2, \hat a^2, \hat v$ by

- $\hat r^2 = \sum_{i=1}^n (y_i - \sigma(\boldsymbol{x}_i^T\hat\beta))^2/n$
- $\hat a^2 = \|\hat\beta\|^2 - \frac{p/n}{(\lambda+\hat v)^2}\hat r^2$ where

$$\begin{cases} \hat v = \sum_{i=1}^n \sigma'(\boldsymbol{x}_i^T\hat\beta)(1-\sigma'(\boldsymbol{x}_i^T\hat\beta)\boldsymbol{x}_i^T\boldsymbol{A}\boldsymbol{x}_i) \\ \boldsymbol{A} = (\sum_{i=1}^n \boldsymbol{x}_i\sigma'(\boldsymbol{x}_i^T\hat\beta)\boldsymbol{x}_i^T)^{-1} \text{ (Hessian)} \end{cases}$$

Approximately normal (e.g., for confidence intervals)

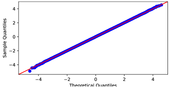
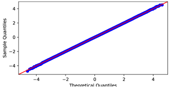
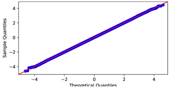
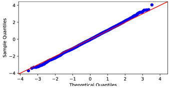
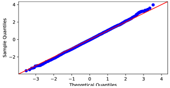
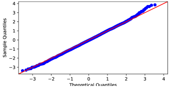
QQplot of $Z_j = \left(\frac{p}{n}\right)^{1/2}\frac{(\hat v+\lambda)}{\hat r}\left(\hat\beta_j - \pm\hat a w_j\right)$ $\begin{cases}\text{shrinking adjustment } \hat a \\ \text{variance adjustment}\end{cases}$

$$\hat{\beta} = \operatorname{argmin}_{\boldsymbol{b}} \frac{1}{n} \sum_{i=1}^{n} (\log(1 + e^{\boldsymbol{x}_i^T \boldsymbol{b}}) - y_i \boldsymbol{x}_i^T \boldsymbol{b}) + \lambda \|\boldsymbol{b}\|^2/2$$

- $\hat{r}^2 = \sum_{i=1}^{n} (y_i - \sigma(\boldsymbol{x}_i^T \hat{\beta}))^2/n$
- $\hat{a}^2 = \|\hat{\beta}\|^2 - \frac{p/n}{(\lambda+\hat{v})^2} \hat{r}^2$ where

$$\begin{cases} \hat{v} = \sum_{i=1}^{n} \sigma'(\boldsymbol{x}_i^T \hat{\beta})(1 - \sigma'(\boldsymbol{x}_i^T \hat{\beta}) \boldsymbol{x}_i^T \boldsymbol{A} \boldsymbol{x}_i) \\ \boldsymbol{A} = (\sum_{i=1}^{n} \boldsymbol{x}_i \sigma'(\boldsymbol{x}_i^T \hat{\beta}) \boldsymbol{x}_i^T)^{-1} \text{ (Here } \sigma(t) = (1 + e^{-t})^{-1} \text{ sigmoid)} \end{cases}$$

- $Z_j = \hat{r}^{-1} \sqrt{p/n} (\hat{v} + \lambda)(\hat{\beta}_j - \pm \hat{a} w_j)$

λ	0.01	0.10	1.00
$\hat{a}^2$	0.630±0.167	0.170±0.039	0.016±0.003
$a_*^2 = (\boldsymbol{w}^T \hat{\beta})^2$	0.610±0.039	0.164±0.009	0.016±0.0009
Z_j for $j : w_j = 0$			
Z_j for $j : w_j \neq 0$			

Logistic Lasso with q repeated measurements

$\forall i \in [n]$ observe $(Y_i^k)_{k=1,\dots,q}$ iid $P(Y_i^k = 1|\boldsymbol{x}_i) = \text{sigmoid}(\boldsymbol{x}_i^T \boldsymbol{\beta}^*)$

$$\hat{\boldsymbol{\beta}} = \min_{\boldsymbol{b}} \sum_{i=1}^n \sum_{k=1}^q \left[\log(1 + e^{\boldsymbol{x}_i^T \boldsymbol{b}}) - Y_i^q \boldsymbol{x}_i^T \boldsymbol{b}\right] + \lambda\sqrt{n}\|\boldsymbol{b}\|_1.$$

Estimate/maximize correlation $\|\hat{\boldsymbol{\beta}}\|^{-1}\hat{\boldsymbol{\beta}}^T\boldsymbol{\beta}^*\|\hat{\boldsymbol{\beta}}^*\|^{-1}$ over λ

Logistic Lasso with q repeated measurements

$\forall i \in [n]$ observe $(Y_i^k)_{k=1,\dots,q}$ iid $P(Y_i^k = 1 | \boldsymbol{x}_i) = \text{sigmoid}(\boldsymbol{x}_i^T \boldsymbol{\beta}^*)$

$$\hat{\boldsymbol{\beta}} = \min_{\boldsymbol{b}} \sum_{i=1}^n \sum_{k=1}^q \left[\log(1 + e^{\boldsymbol{x}_i^T \boldsymbol{b}}) - Y_i^q \boldsymbol{x}_i^T \boldsymbol{b} \right] + \lambda \sqrt{n} \|\boldsymbol{b}\|_1.$$

Estimate/maximize correlation $\|\hat{\boldsymbol{\beta}}\|^{-1} \hat{\boldsymbol{\beta}}^T \boldsymbol{\beta}^* \|\hat{\boldsymbol{\beta}}^*\|^{-1}$ over λ

Define Vector $\hat{\boldsymbol{\psi}} \in \mathbb{R}^n$ has components $\hat{\psi}_i = -\sum_{k=1}^q \ell'(\boldsymbol{x}_i^T \hat{\boldsymbol{\beta}};\ Y_i^k)$

$$\frac{\hat{\boldsymbol{\beta}}^T \boldsymbol{\beta}^*}{\|\hat{\boldsymbol{\beta}}^*\|} \approx \hat{a} := \frac{\left(\frac{\hat{v}}{n} \|\boldsymbol{X}\hat{\boldsymbol{\beta}} - \hat{\gamma}\hat{\boldsymbol{\psi}}\|^2 + \frac{1}{n}\hat{\boldsymbol{\psi}}^\top \boldsymbol{X}\hat{\boldsymbol{\beta}} - \hat{\gamma}\hat{r}^2 \right)^2}{\frac{1}{n^2}\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}^T\hat{\boldsymbol{\psi}}\|^2 + \frac{2\hat{v}}{n}\hat{\boldsymbol{\psi}}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} + \frac{\hat{v}^2}{n}\|\boldsymbol{X}\hat{\boldsymbol{\beta}} - \hat{\gamma}\hat{\boldsymbol{\psi}}\|^2 - \frac{p}{n}\hat{r}^2}.$$

Logistic Lasso with q repeated measurements

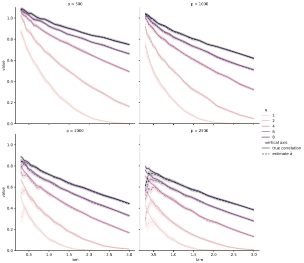
$\forall i \in [n]$ observe $(Y_i^k)_{k=1,\dots,q}$ iid $P(Y_i^k = 1|\boldsymbol{x}_i) = \text{sigmoid}(\boldsymbol{x}_i^T \boldsymbol{\beta}^*)$

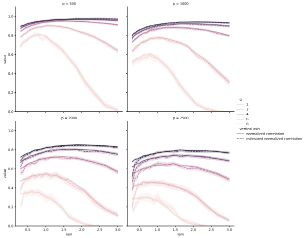
$$\hat{\boldsymbol{\beta}} = \min_{\boldsymbol{b}} \textstyle\sum_{i=1}^n \sum_{k=1}^q \left[\log(1 + e^{\boldsymbol{x}_i^T \boldsymbol{b}}) - Y_i^q \boldsymbol{x}_i^T \boldsymbol{b}\right] + \lambda\sqrt{n}\|\boldsymbol{b}\|_1.$$

Estimate/maximize correlation $\|\hat{\boldsymbol{\beta}}\|^{-1}\hat{\boldsymbol{\beta}}^T\boldsymbol{\beta}^*\|\hat{\boldsymbol{\beta}}^*\|^{-1}$ over λ

Define Vector $\hat{\boldsymbol{\psi}} \in \mathbb{R}^n$ has components $\hat{\psi}_i = -\sum_{k=1}^q \ell'(\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}};\ Y_i^k)$

$$\frac{\hat{\boldsymbol{\beta}}^T\boldsymbol{\beta}^*}{\|\hat{\boldsymbol{\beta}}^*\|} \approx \hat{a} := \frac{\left(\frac{\hat{v}}{n}\|\boldsymbol{X}\hat{\boldsymbol{\beta}} - \hat{\gamma}\hat{\boldsymbol{\psi}}\|^2 + \frac{1}{n}\hat{\boldsymbol{\psi}}^\top\boldsymbol{X}\hat{\boldsymbol{\beta}} - \hat{\gamma}\hat{r}^2\right)^2}{\frac{1}{n^2}\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}^T\hat{\boldsymbol{\psi}}\|^2 + \frac{2\hat{v}}{n}\hat{\boldsymbol{\psi}}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} + \frac{\hat{v}^2}{n}\|\boldsymbol{X}\hat{\boldsymbol{\beta}} - \hat{\gamma}\hat{\boldsymbol{\psi}}\|^2 - \frac{p}{n}\hat{r}^2}.$$

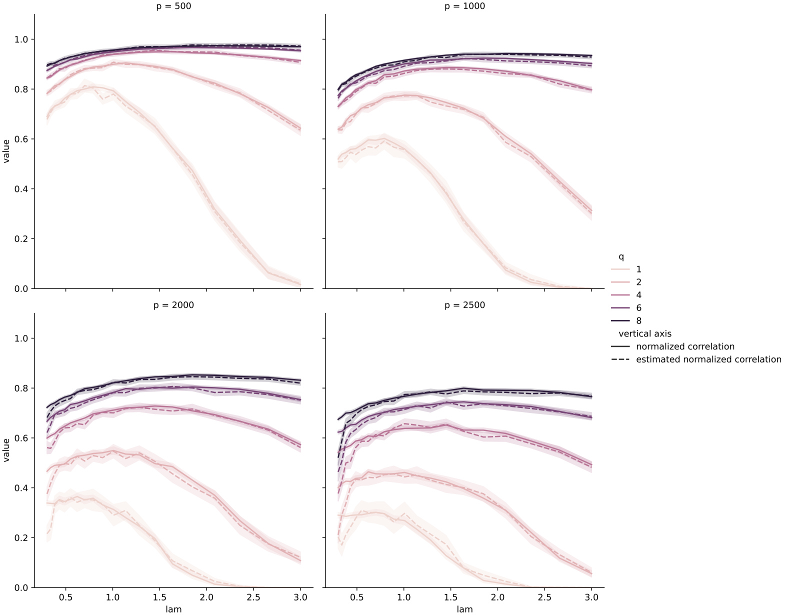




$\hat{\boldsymbol{\beta}}^T\boldsymbol{\beta}^*$

$$\frac{\hat{\boldsymbol{\beta}}^T\boldsymbol{\beta}^*}{\|\hat{\boldsymbol{\beta}}\|\|\boldsymbol{\beta}^*\|}$$

Logistic Lasso with repeated measurements: $\frac{\hat{\beta}^T \beta^*}{\|\hat{\beta}\| \|\beta^*\|}$



Literature on generalized linear models (linear, logistic, ...)

Regime: $n/p \to \delta$ (=constant)

M-estimator with separable penalty

$$\hat{\beta}(\boldsymbol{y}, \boldsymbol{X}) = \underset{\boldsymbol{b}\in\mathbb{R}^p}{\text{argmin}} \frac{1}{n}\sum_{i=1}^{n} \ell_{y_i}\left(\boldsymbol{x}_i^T \boldsymbol{b}\right) + \frac{1}{p}\sum_{j=1}^{p} \tilde{f}(b_j)$$

Informal result:

If $\boldsymbol{X}$ has iid $N(0, \frac{1}{p})$ entries, Then the empirical distribution of $(\hat{\beta}_j)_{j=1,\dots,p}$ is approx. the same as the empirical distribution of

$$\text{prox}\left[\bar{\gamma}\ \tilde{f}\right]\left(\bar{c}\ \beta_j^* + \bar{c}'\ Z_j\right), \qquad Z_j \sim N(0,1)$$

for some constants $\bar{\gamma}, \bar{c}, \bar{c}'$ depending on $\delta = \lim \frac{n}{p}$, the penalty $\tilde{f}$, the data-generating process and loss function.

Why find $\bar{\gamma}, \bar{c}, \bar{c}'$?

$\bar{\gamma}, \bar{c}, \bar{c}'$ characterize MSE $\frac{1}{p}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|^2$, correlation $\frac{1}{p}\hat{\boldsymbol{\beta}}^T\boldsymbol{\beta}^*$, etc

How to find $\bar{\gamma}, \bar{c}, \bar{c}'$?

Some literature in linear models

El Karoui et al (2013), Donoho and Montanari (2016)
$\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{b}\in\mathbb{R}^p} \sum_{i=1}^n \mathcal{L}(y_i - \boldsymbol{x}_i^T\boldsymbol{b})$ for some convex $\mathcal{L} : \mathbb{R} \to \mathbb{R}$.

System of two equations

$$\begin{cases} \delta^{-1}\sigma^2 = \mathbb{E}\big[(\text{prox}[\gamma\mathcal{L}](\varepsilon_1 + \sigma Z) - \varepsilon_1 - \sigma Z)^2\big], \\ 1 - \delta^{-1} = \mathbb{E}\big[\text{prox}[\gamma\mathcal{L}]'(\varepsilon_1 + \sigma Z)\big], \end{cases}$$

with two unknowns (σ, γ), where $Z \sim N(0,1)$ is independent of ε_1.

If $\boldsymbol{X}$ has iid entries then $\|\hat{\boldsymbol{\beta}}\|^2 \to^P \sigma^2$
Also, asymptotic normality results for $\hat{\beta}_j$

Similar work for the Lasso and $\boldsymbol{X}$ with iid $N(0,1)$ entries
(Bayati and Montanari 2011)

Logistic Regression (Sur and Candes 2018)

- Logistic model, $\rho'(u) = 1/(1+e^{-u})$ is the sigmoid
- $\boldsymbol{\beta}^*$ iid entries with law β and $\mathbb{E}[\beta^2] = \kappa^2$
- $\boldsymbol{x}_i \sim N(0, \frac{1}{p}\boldsymbol{I}_p)$ ($\boldsymbol{\Sigma} = \frac{1}{p}\boldsymbol{I}_p$)
- $n, p \to \infty$ with $n/p \to \delta$.

System with three unknowns σ, α, γ

$$\begin{cases} \delta^{-1}\sigma^2 = 2\mathbb{E}\left[\rho'(-\kappa Z_1)\left(\gamma\rho'(\text{prox}[\gamma\rho](\kappa\alpha Z_1 + \sigma Z_2))\right)^2\right], \\ 0 = 2\mathbb{E}\left[\rho'(-\kappa Z_1) Z_1 \gamma\rho'(\text{prox}[\gamma\rho](\kappa\alpha Z_1 + \sigma Z_2))\right], \\ 1 - \delta^{-1} = 2\mathbb{E}\left[\rho'(-\kappa Z_1)\text{prox}[\gamma\rho]'(\kappa\alpha Z_1 + \sigma Z_2)\right]. \end{cases}$$

With $(\bar{\alpha}, \bar{\sigma}, \bar{\gamma})$ denoting the solution

$$\frac{1}{p}\sum_{j=1}^{p} \phi\left(\hat{\beta}_j - \bar{\alpha}\beta_j^*, \beta_j^*\right) \to^{\mathbb{P}} \mathbb{E}\left[\phi\left(\bar{\sigma}Z, \beta\right)\right]$$

where $Z \sim N(0,1)$ is independent of β,

Logistic loss+penalty $g(\boldsymbol{b}) = \sum_{j=1}^{p} \frac{\tilde{f}(b_j)}{p}$ (Salehi et al 2019)

- Logistic model
- $\boldsymbol{\beta}^*$ iid entries with law β and $\mathbb{E}[\beta^2] = \kappa^2$
- $\boldsymbol{x}_i \sim N(0, \frac{1}{p}\boldsymbol{I}_p)$ ($\boldsymbol{\Sigma} = \frac{1}{p}\boldsymbol{I}_p$)

System with six unknowns $(\alpha, \sigma, \gamma, \theta, \tau, r)$,

$$\begin{cases} \kappa^2\alpha = \mathbb{E}\big[\beta \text{prox}[\sigma\tau\tilde{f}(\cdot)](\sigma\tau(\theta\beta + \delta^{-1/2} rZ))\big], \\ \sqrt{\delta} r\gamma = \mathbb{E}\big[Z \text{prox}[\sigma\tau\tilde{f}(\cdot)](\sigma\tau(\theta\beta + \delta^{-1/2} rZ))\big], \\ \kappa^2\alpha^2 + \sigma^2 = \mathbb{E}\big[\{\text{prox}[\sigma\tau\tilde{f}(\cdot)](\sigma\tau(\theta\beta + \delta^{-1/2} rZ))\}^2\big], \\ r^2\gamma^2 = 2\mathbb{E}\big[\rho'(-\kappa Z_1)(\kappa\alpha Z_1 + \sigma Z_2 - \text{prox}[\gamma\rho](\kappa\alpha Z_1 + \sigma Z_2))^2\big], \\ -\theta\gamma = 2\mathbb{E}\big[\rho''(-\kappa Z_1)\text{prox}[\gamma\rho](\kappa\alpha Z_1 + \sigma Z_2)\big], \\ 1 - \gamma/(\sigma\tau) = 2\mathbb{E}\big[\rho'(-\kappa Z_1)\text{prox}[\gamma\rho]'(\kappa\alpha Z_1 + \sigma Z_2)\big] \end{cases}$$

Logistic loss+penalty $g(\boldsymbol{b}) = \sum_{j=1}^{p} \frac{\tilde{f}(b_j)}{p}$ (Salehi et al 2019)

- Logistic model
- $\boldsymbol{\beta}^*$ iid entries with law β and $\mathbb{E}[\beta^2] = \kappa^2$
- $\boldsymbol{x}_i \sim N(0, \frac{1}{p}\boldsymbol{I}_p)$ $(\boldsymbol{\Sigma} = \frac{1}{p}\boldsymbol{I}_p)$
- $n, p \to \infty$ with $n/p \to \delta$.

System with six unknowns $(\alpha, \sigma, \gamma, \theta, \tau, r)$,

If unique solution $(\bar{\alpha}, \bar{\sigma}, \bar{\gamma}, \bar{\theta}, \bar{\tau}, \bar{r})$ then for any locally Lipschitz Φ

$$\frac{1}{p}\sum_{j=1}^{p} \Phi\Big(\hat{\beta}_j, \beta_j^*\Big) \to^{\mathbb{P}} \mathbb{E}\Big[\Phi\Big(\mathrm{prox}[\bar{\sigma}\bar{\tau}\tilde{f}(\cdot)]\big(\bar{\sigma}\bar{\tau}(\bar{\theta}\beta + \delta^{-1/2}\bar{r}Z)\big), \beta\Big)\Big]$$

See Loureiro et al. (2021) for a unifying theory. Informally:

$$\hat{\beta}_j \approx \mathrm{prox}[\bar{\sigma}\bar{\tau}\tilde{f}(\cdot)]\big(\bar{\sigma}\bar{\tau}(\bar{\theta}\beta_j^* + \delta^{-1/2}\bar{r}Z_j)\big), \qquad [\text{where } Z_j \sim N(0,1)]$$

A peek at the results (informal)

Single index model

iid $(\boldsymbol{x}_i, y_i)_{i=1,\dots,n}$ with Gaussian $\boldsymbol{x}_i \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$ and

$$y_i = F(\boldsymbol{x}_i^T \boldsymbol{w}, U_i), \qquad \mathrm{Var}[\boldsymbol{x}_i^T \boldsymbol{w}] = \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{w}\|^2 = 1.$$

M-estimator (in this slide, with separable penalty)

$$\hat{\beta}(\boldsymbol{y}, \boldsymbol{X}) = \mathrm{argmin}_{\boldsymbol{b}\in\mathbb{R}^p} \tfrac{1}{n}\textstyle\sum_{i=1}^n \ell_{y_i}\big(\boldsymbol{x}_i^T \boldsymbol{b}\big) + \tfrac{1}{p}\textstyle\sum_{j=1}^p \tilde{f}(b_j)$$

Result: empirical distribution $(\hat{\beta}_j)_{j=1,\dots,p}$ well-approximated as

$$\hat{\beta}_j \approx \mathrm{prox}\Big[\frac{1}{\hat{v}}\tilde{f}\Big]\Big(\pm w_j \frac{\hat{t}}{\hat{v}} + \frac{1}{\sqrt{\delta}}\frac{\hat{r}}{\hat{v}} Z_j\Big), \qquad \text{where } Z_j \sim N(0,1)$$

- $\pm w_j$ the j-th entry of the index $\boldsymbol{w}$ up to an unidentifiable $\pm$.
- $(\hat{v}, \hat{t}, \hat{r})$ are **observable** scalars
- Why find $(\hat{v}, \hat{t}, \hat{r})$? Confidence interval, $\widehat{MSE}$, $\widehat{correlation}$
- How to find $(\hat{v}, \hat{t}, \hat{r})$?

Derivatives: for some matrix $\hat{\boldsymbol{A}} \in \mathbb{R}^{p \times p}$, with $\hat{\psi}_i = -\ell_{y_i}(\boldsymbol{x}_i^T \hat{\boldsymbol{\beta}})$,

$$\frac{\partial}{\partial x_{ij}} \hat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X}) = \hat{\boldsymbol{A}} \boldsymbol{e}_j \hat{\psi}_i - \hat{\boldsymbol{A}} \boldsymbol{X}^T \boldsymbol{D} \boldsymbol{e}_i \hat{\beta}_j, \qquad \boldsymbol{D} = \operatorname{diag}(\ell''_{y_i}(\boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}))$$

Notation $\boldsymbol{V} = \boldsymbol{D} - \boldsymbol{D} \boldsymbol{X} \hat{\boldsymbol{A}} \boldsymbol{X}^T \boldsymbol{D}$ (matrix $n \times n$), $\hat{\text{df}} = \operatorname{Tr}[\boldsymbol{X} \hat{\boldsymbol{A}} \boldsymbol{X}^T \boldsymbol{D}]$

Derivatives: for some matrix $\hat{\boldsymbol{A}} \in \mathbb{R}^{p\times p}$, with $\hat{\psi}_i = -\ell_{y_i}(\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})$,

$$\frac{\partial}{\partial x_{ij}}\hat{\boldsymbol{\beta}}(\boldsymbol{y},\boldsymbol{X}) = \hat{\boldsymbol{A}}\boldsymbol{e}_j\hat{\psi}_i - \hat{\boldsymbol{A}}\boldsymbol{X}^T\boldsymbol{D}\boldsymbol{e}_i\hat{\beta}_j, \qquad \boldsymbol{D} = \operatorname{diag}(\ell''_{y_i}(\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}))$$

Notation $\boldsymbol{V} = \boldsymbol{D} - \boldsymbol{D}\boldsymbol{X}\hat{\boldsymbol{A}}\boldsymbol{X}^T\boldsymbol{D}$ (matrix $n\times n$), $\hat{\mathrm{df}} = \operatorname{Tr}[\boldsymbol{X}\hat{\boldsymbol{A}}\boldsymbol{X}^T\boldsymbol{D}]$

$(\hat{v},\hat{t},\hat{r})$ used to describe the empirical dist. of $(\hat{\beta}_j)_{j=1,\dots,p}$

The three others $\hat{\gamma},\hat{a}^2,\hat{\sigma}^2$ for the empirical dist. of $(\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})_{i=1,\dots,n}$.

$$\begin{cases} \hat{v} \stackrel{\text{def}}{=} \frac{1}{n}\operatorname{Tr}[\boldsymbol{V}], \\ \hat{r} \stackrel{\text{def}}{=} (\frac{1}{n}\|\hat{\boldsymbol{\psi}}\|^2)^{1/2} = (\frac{1}{n}\sum_{i=1}^n \ell'_{y_i}(\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})^2)^{1/2}, \\ \hat{t}^2 \stackrel{\text{def}}{=} \frac{1}{n^2}\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}^T\hat{\boldsymbol{\psi}}\|^2 + \frac{2\hat{v}}{n}\hat{\boldsymbol{\psi}}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} + \frac{\hat{v}^2}{n}\|\boldsymbol{X}\hat{\boldsymbol{\beta}} - \hat{\gamma}\hat{\boldsymbol{\psi}}\|^2 - \frac{p}{n}\hat{r}^2, \\ \hat{\gamma} \stackrel{\text{def}}{=} \frac{\hat{\mathrm{df}}}{n\hat{v}} = \frac{\hat{\mathrm{df}}}{\operatorname{Tr}[\boldsymbol{V}]}, \\ \hat{a}^2 \stackrel{\text{def}}{=} \hat{t}^{-2}\left(\frac{\hat{v}}{n}\|\boldsymbol{X}\hat{\boldsymbol{\beta}} - \hat{\gamma}\hat{\boldsymbol{\psi}}\|^2 + \frac{1}{n}\hat{\boldsymbol{\psi}}^\top\boldsymbol{X}\hat{\boldsymbol{\beta}} - \hat{\gamma}\hat{r}^2\right)^2, \\ \hat{\sigma}^2 \stackrel{\text{def}}{=} \frac{1}{n}\|\boldsymbol{X}\hat{\boldsymbol{\beta}} - \hat{\gamma}\hat{\boldsymbol{\psi}}\|^2 - \hat{a}^2. \end{cases}$$

Much simpler expressions for special cases

E.g., for unregularized M-estimation ($g = 0$):

$$\frac{\partial}{\partial x_{ij}} \hat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X}) = \hat{\boldsymbol{A}} \boldsymbol{e}_j \hat{\psi}_i - \hat{\boldsymbol{A}} \boldsymbol{X}^T \boldsymbol{D} \boldsymbol{e}_i \hat{\beta}_j, \qquad \hat{\boldsymbol{A}} = \Big(\sum_{i=1}^{n} \boldsymbol{x}_i \ell''_{y_i}(\boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}) \boldsymbol{x}_i^T \Big)^{-1}$$

$$\hat{v} = \frac{1}{n} \sum_{i=1}^{n} \ell''_{y_i}(\boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}) \Big[1 - \ell''_{y_i}(\boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}) \boldsymbol{x}_i^T \hat{\boldsymbol{A}} \boldsymbol{x}_i \Big], \qquad \hat{r}^2 = \frac{1}{n} \sum_{i=1}^{n} \ell'_{y_i}(\boldsymbol{x}_i^T \hat{\boldsymbol{\beta}})^2$$

$$\hat{\text{df}} = p, \qquad \hat{\gamma} = \frac{p/n}{\hat{v}}, \qquad \hat{a}^2 = \frac{\|\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2}{n} - \frac{p}{n}\Big(1 - \frac{p}{n}\Big) \frac{\hat{r}^2}{\hat{v}^2}$$

$$\hat{t}^2 = \hat{a}^2 \hat{v}^2, \qquad \hat{\sigma}^2 = \frac{p}{n} \Big(\frac{\hat{r}}{\hat{v}} \Big)^2.$$

Here, the fact that $\hat{\text{df}} = p$ justifies the notation $\hat{\text{df}}$.

Much simpler expressions for special cases

E.g., for Least-Squares $\ell_{y_i}(u) = \frac{1}{2}(u - y_i)^2$, penalty $g = 0$:

$$\frac{\partial}{\partial x_{ij}} \hat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X}) = \hat{\boldsymbol{A}} \boldsymbol{e}_j \hat{\psi}_i - \hat{\boldsymbol{A}} \boldsymbol{X}^T \boldsymbol{D} \boldsymbol{e}_i \hat{\beta}_j, \qquad \hat{\boldsymbol{A}} = \left(\boldsymbol{X}^T \boldsymbol{X} \right)^{-1}$$

$$\hat{v} = 1 - p/n, \qquad \hat{r}^2 = \frac{1}{n} \|\boldsymbol{y} - \boldsymbol{X} \hat{\boldsymbol{\beta}}\|^2$$

$$\hat{\mathrm{df}} = p, \qquad \hat{\gamma} = \frac{p/n}{\hat{v}}, \qquad \hat{a}^2 = \frac{\|\boldsymbol{X} \hat{\boldsymbol{\beta}}\|^2}{n} - \frac{p}{n} \left(1 - \frac{p}{n}\right) \frac{\|\boldsymbol{y} - \boldsymbol{X} \hat{\boldsymbol{\beta}}\|^2 / n}{\hat{v}^2}$$

$$\hat{t}^2 = \hat{a}^2 \hat{v}^2, \qquad \hat{\sigma}^2 = \frac{p}{n} \left(\frac{\hat{r}}{\hat{v}} \right)^2.$$

Here, the fact that $\hat{\mathrm{df}} = p$ justifies the notation $\hat{\mathrm{df}}$.

Theorem 4.1

Assumptions:

- $\boldsymbol{x}_i \sim N(0, \boldsymbol{\Sigma})$, condition number of Σ bounded by κ
- $1000 \geq n/p \geq \delta$
- penalty τ-strongly convex
- $\hat{\beta}_j^{(d)} = \hat{\beta}_j + \mathrm{Tr}[\boldsymbol{V}]^{-1} \boldsymbol{e}_j^T \boldsymbol{\Sigma}^{-1} \boldsymbol{X}^T \hat{\psi}, \qquad \Omega_{jj} = (\boldsymbol{\Sigma}^{-1})_{jj}$

Then for all $j = 1, ..., p$, there exists $Z_j \sim N(0,1)$ such that

$$\frac{1}{p} \sum_{j=1}^{p} \mathbb{E}\Big[\Big(\frac{\sqrt{n}}{\Omega_{jj}^{1/2}}\Big(\frac{\hat{v}}{\hat{r}} \hat{\beta}_j^{(d)} - \frac{\pm \hat{t}}{\hat{r}} w_j\Big) - Z_j\Big)^2\Big] \leq \frac{C_1(\delta, \tau, \kappa)}{\sqrt{p}}$$

where $\pm$ denotes an unidentifiable random sign.

Theorem 4.1

Assumptions:

- $\boldsymbol{x}_i \sim N(0, \boldsymbol{\Sigma})$, condition number of Σ bounded by κ
- $1000 \geq n/p \geq \delta$
- penalty τ-strongly convex
- $\hat{\beta}_j^{(d)} = \hat{\beta}_j + \text{Tr}[\boldsymbol{V}]^{-1} \boldsymbol{e}_j^T \boldsymbol{\Sigma}^{-1} \boldsymbol{X}^T \hat{\boldsymbol{\psi}}, \qquad \Omega_{jj} = (\boldsymbol{\Sigma}^{-1})_{jj}$

Then for all $j = 1, ..., p$, there exists $Z_j \sim N(0, 1)$ such that

$$\frac{1}{p} \sum_{j=1}^{p} \mathbb{E}\Big[\Big(\frac{\sqrt{n}}{\Omega_{jj}^{1/2}}\Big(\frac{\hat{v}}{\hat{r}} \hat{\beta}_j^{(d)} - \frac{\pm \hat{t}}{\hat{r}} w_j\Big) - Z_j\Big)^2\Big] \leq \frac{C_2(\delta, \tau, \kappa)}{\sqrt{p}}$$

where $\pm$ denotes an unidentifiable random sign.

- Consequence of Theorem 4.1: proximal representation for $\hat{\boldsymbol{\beta}}$
  $\hat{\beta}_j \approx \text{prox}\Big[\frac{1}{\hat{v}} \tilde{f}\Big]\Big(\pm w_j \frac{\hat{t}}{\hat{v}} + \frac{1}{\sqrt{\delta}} \frac{\hat{r}}{\hat{v}} Z_j\Big)$ for sep. penalty, $\boldsymbol{\Sigma} = \frac{1}{p} \boldsymbol{I}_p$
- Theorem 4.3: Proximal representation for $\boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}$
- Theorem 4.4: correlation estimation $\hat{a}^2 \approx (\boldsymbol{w}^T \hat{\boldsymbol{\beta}})^2$

Take home

- Empirical distribution $\hat\beta_j \approx \text{prox}\left[\frac{1}{\hat v}\tilde f\right]\left(\pm w_j \frac{\hat t}{\hat v} + \frac{1}{\sqrt{\delta}}\frac{\hat r}{\hat v} Z_j\right)$ and confidence intervals for the entries $\pm w_j$ of the index $\boldsymbol{w}$
- Data-driven parameters in the proximal representation can be read in the derivatives of $\hat{\boldsymbol\beta}(\boldsymbol y, \boldsymbol X)$ with respect to $\boldsymbol X$,

$$\frac{\partial}{\partial x_{ij}}\hat{\boldsymbol\beta}(\boldsymbol y, \boldsymbol X) = \hat{\boldsymbol A}\boldsymbol e_j \hat\psi_i - \hat{\boldsymbol A}\boldsymbol X^T \boldsymbol D \boldsymbol e_i \hat\beta_j,$$

$\hat v = \frac{1}{n}\text{Tr}[\boldsymbol D - \boldsymbol D \boldsymbol X \hat{\boldsymbol A} \boldsymbol X^T \boldsymbol D]$ where $\boldsymbol D = \text{diag}(\ell''_{y_i}(\boldsymbol x_i^T \hat{\boldsymbol\beta}))$.

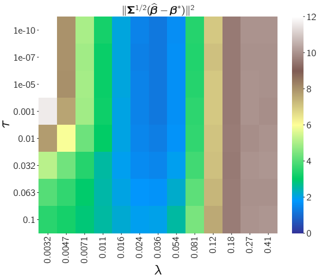
- Without solving the deterministic fixed-point equations obtained by Approximate Message Passing or Gordon's CGMT

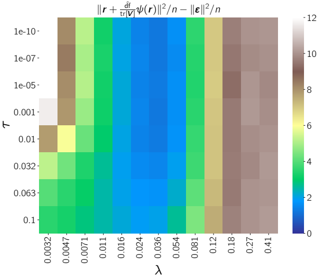
Linear models: Estimating Generalization/param. tuning

$$\hat{\beta}(\boldsymbol{y}, \boldsymbol{X}) = \operatorname*{argmin}_{\boldsymbol{b}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{x}_i^\top \boldsymbol{b} - y_i) + \lambda\|\boldsymbol{b}\|_1 + \tau\|\boldsymbol{b}\|^2/2$$

Huber Loss $\ell(u) = \int_0^{|u|} \min(1, t)dt$ with Elastic-Net penalty

Two tuning parameters (λ, τ) in the Elastic-Net penalty



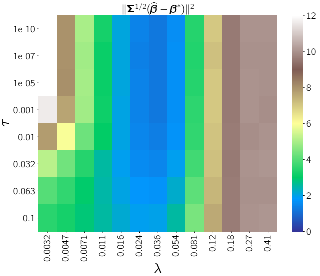


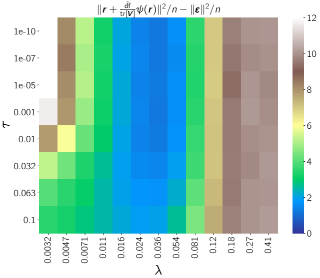
Linear models: Estimating Generalization/param. tuning

$$\hat{\beta}(\boldsymbol{y}, \boldsymbol{X}) = \underset{\boldsymbol{b}\in\mathbb{R}^p}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{x}_i^\top \boldsymbol{b} - y_i) + \lambda\|\boldsymbol{b}\|_1 + \tau\|\boldsymbol{b}\|^2/2$$

Huber Loss $\ell(u) = \int_0^{|u|} \min(1, t)dt$ with Elastic-Net penalty

Two tuning parameters (λ, τ) in the Elastic-Net penalty



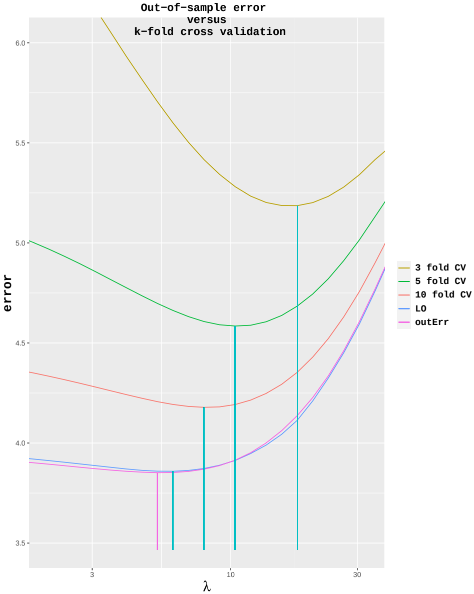


With $\hat{\mathrm{df}} = \mathrm{Tr}[\boldsymbol{X}\hat{\boldsymbol{A}}\boldsymbol{X}^T\boldsymbol{D}]$, $\hat{v} = \mathrm{Tr}[\boldsymbol{D}] - \hat{\mathrm{df}}/n$, Theory gives approx.:

$$\|\boldsymbol{\Sigma}^{1/2}(\hat{\beta} - \beta^*)\|^2 + \frac{\|\boldsymbol{\varepsilon}\|^2}{n} \approx \frac{1}{n}\left\|(\boldsymbol{y} - \boldsymbol{X}\hat{\beta}) + \frac{\hat{\mathrm{df}}/n}{\hat{v}}\ell'(\boldsymbol{y} - \boldsymbol{X}\hat{\beta})\right\|^2$$

K-Fold Cross-validation suffers sample-size bias

Figure 1 from *Consistent Risk Estimation in Moderately High-Dimensional Linear Regression* by Xu, Maleki, Rad, Hsu (arXiv:1902.01753)



References

Approximate Message Passing (AMP)

- *The LASSO risk for Gaussian matrices* (Bayati and Montanari 2011)
- *A unifying tutorial on Approximate Message Passing*, (Feng, Venkataramanan, Rush, Samworth, 2021)
- Logistic Regression (Sur and Candes 2018)

Gordon's Convex Gaussian Min-Max Theorem

- *Precise Error Analysis of Regularized M-estimators in High-dimensions* Thrampoulidis et al (2015)
- Reguarlized logistic regression: Salehi, Abbasi Hassibi (2019)
- Lasso: Miolane, Montanari (2018)
- Lasso, correlated $\boldsymbol{X}$: Celentano, Montanari and Wei (2021)
- *Learning curves of generic features maps for realistic datasets with a teacher-student model* (Loureiro et al 2021)

This work and related techniques

- Linear model: *Out-of-sample error estimate for robust M-estimators with convex penalty* (B, 2020)
- Linear model: *Asymptotic normality of robust M-estimators with convex penalty*, (Bellec, Shen, Zhang 2021)
- Linear model: *Derivatives and residual distribution of regularized M-estimators with application to adaptive tuning*, (B and Shen 2021)
- Single-index: *Observable adjustments in single-index models for regularized M-estimators*, (B, 2022)