Uniform Linear Representation for Post-selection Inference[1]

Arun Kumar Kuchibhotla

Department of Statistics
University of Pennsylvania

July 12, 2018

Workshop Model Selection, Regularization, and Inference

[1]Joint work with "Larry's Group" at Wharton, including Larry Brown, Andreas Buja, Edward George, Linda Zhao and Junhui Cai.

In Memory of

LAWRENCE D. BROWN †
1940 – 2018.

Outline

Outline

Some History

- The practice of data analysis often involves exploring the data thoroughly before a formal modeling begins. Exploratory Data Analysis (EDA) is an example.
- Reproducibility/replicability crisis in science is sometimes attributed to this type of data analysis.
- Another reason for invalid statistical inference is the “blind” use of classical tools as if all models used are correctly specified.

Wanted: Valid Inference under Possible Misspecification and Arbitrary data-driven Modeling!

Some Review

- Valid inference under data-driven modeling is the current “hot topic”: Post-selection Inference (PoSI).
- Berk et al. (2013) solved PoSI in a well-specified linear regression.
- Jonathan Taylor and others have developed selective inference techniques: Lee et al. (2016), Tibshirani et al. (2016), Tian et al. (2016), for example.
- Because of various issues like *Fragility* and, *impossibility results of Leeb and Postscher (2008)*, related to selective inference, we favor the ideology of Berk et al. (2013).
- Bachoc et al. (2016) generalized the setting of Berk et al. (2013) to allow misspecification but deals fixed number of models.

Motivating Example

Suppose $(X_i, Y_i) \in \mathbb{R}^{p+1}$ are independent random vectors and select a subset of variables $\hat{M} \subseteq \{1, 2, \ldots, p\}$, using any subset selection procedure. Compute the OLS linear regression estimator $\hat{\beta}_{n,\hat{M}}$:

$$\hat{\beta}_{n,\hat{M}} := \underset{\theta \in \mathbb{R}^{|\hat{M}|}}{\arg\min} \; \frac{1}{n} \sum_{i=1}^{n} \left\{ Y_i - \theta^\top X_i(\hat{M}) \right\}^2 .$$

Here $X_i(\hat{M})$ represents the sub-vector of X_i with indices in $\hat{M}$.

Problem

– *What is $\hat{\beta}_{n,\hat{M}}$ estimating?*

– *How to perform inference for the resulting target?*

– *How large can $|\hat{M}|$ be in terms of the sample size n?*

Some Notation

- Define set of models of size bounded by k as

$$\mathcal{M}(k) := \{M \subseteq \{1, 2, \ldots, p\} : 1 \le |M| \le k\}.$$

- Define the Gram matrix and covariance vector for model M as

$$\hat{\Sigma}_n := \hat{\mathbb{E}}_n \left[X_i X_i^\top \right], \quad \text{and} \quad \hat{\Gamma}_n := \hat{\mathbb{E}}_n \left[X_i Y_i \right],$$
$$\Sigma_n := \bar{\mathbb{E}}_n \left[X_i X_i^\top \right], \quad \text{and} \quad \Gamma_n := \bar{\mathbb{E}}_n \left[X_i Y_i \right].$$

Here $\hat{\mathbb{E}}_n[\cdot] = \sum_{i=1}^n [\cdot]/n$ and $\bar{\mathbb{E}}_n[\cdot] = \sum_{i=1}^n \mathbb{E}[\cdot]/n$.

- So, the OLS estimate $\hat{\beta}_{n,M}$ and target $\beta_{n,M}$ for model M are

$$\hat{\beta}_{n,M} = \left(\hat{\Sigma}_n(M) \right)^{-1} \hat{\Gamma}_n(M), \quad \beta_{n,M} := \left(\Sigma_n(M) \right)^{-1} \Gamma_n(M).$$

- For $1 \le j \le |M|$, let $\hat{\beta}_{n,M}(j)$ represent the j-th coordinate of $\hat{\beta}_{n,M}$.

Uniform-in-model Result (Kuchibhotla et al. 2018b)

If the observations $(X_i, Y_i) \in \mathbb{R}^{p+1}$ are independent and sub-Gaussian, then

$$\max_{|M|\le k} \left\|\hat{\beta}_{n,M} - \beta_{n,M}\right\|_2 = O_p\left(\sqrt{\frac{k\log(ep/k)}{n}}\right),$$

and so,

$$\max_{j\in M, |M|\le k} \left|\hat{\beta}_{n,M}(j) - \beta_{n,M}(j)\right| = O_p\left(\sqrt{\frac{k\log(ep/k)}{n}}\right).$$

Trivial Inequality: If $|\hat{M}| \le k$, then

$$\left\|\hat{\beta}_{n,\hat{M}} - \beta_{n,\hat{M}}\right\|_2 \le \max_{|M|\le k} \left\|\hat{\beta}_{n,M} - \beta_{n,M}\right\|_2 = O_p\left(\sqrt{\frac{k\log(ep/k)}{n}}\right).$$

So, $\hat{\beta}_{n,\hat{M}}$ is "estimating" $\beta_{n,\hat{M}}$.

Outline

General Framework

- Let $\{\theta_q : q \in \mathcal{Q}\}$ be a collection of real-valued functionals.
- Let $\{\hat{\theta}_q : q \in \mathcal{Q}\}$ be a collection of estimators from independent random variables $Z_1, \ldots, Z_n$ satisfying for functions $\{\psi_{n,q}(\cdot)\}$,

$$\sqrt{n}\left(\hat{\theta}_q - \theta_q\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_{n,q}(Z_i) + R_{n,q} \qquad \text{(AULR)}$$

  (**Asymptotic Uniform Linear Representation**).
- The analogy to model selection:
  - $\mathcal{Q} = \{(j, M) : j \in M, M \in \mathcal{M}(k)\}$.
  - For $q = (j, M)$, $\theta_q = \beta_{n,M}(j)$ (functional) and $\hat{\theta}_q = \hat{\beta}_{n,M}(j)$ (estimator).
  - Inference is sought for $\theta_{\hat{q}}$ (a random functional).

The PoSI Problem

- If $R_{n,q}$ is uniformly close to zero, then $\hat{\theta}_{\hat{q}} - \theta_{\hat{q}} \approx 0$. Without further information, a non-random target for $\hat{\theta}_{\hat{q}}$ does not exist.
- The PoSI problem (Confidence Regions Version) is as follows:

Problem

Construct a set of confidence regions (depending on α)

$$\{\hat{\mathcal{R}}_{n,q} : \; q \in \mathcal{Q}\}$$

for some non-random set of models, $\mathcal{Q}$, such that for any random model $\hat{q}$ with $\mathbb{P}(\hat{q} \in \mathcal{Q}) = 1$,

$$\liminf_{n\to\infty} \; \mathbb{P}\left(\theta_{\hat{q}} \in \hat{\mathcal{R}}_{n,\hat{q}}\right) \geq 1 - \alpha.$$

Equivalence of PoSI and Simultaneous Inference

Theorem (Kuchibhotla et al. (2018a))

For any set of confidence regions $\{\hat{\mathcal{R}}_{n,q} : q \in \mathcal{Q}\}$ and $\alpha \in [0,1]$, the following are EQUIVALENT:

1. *the post-selection inference problem is solved, that is,*

$$\mathbb{P}\left(\theta_{\hat{q}} \in \hat{\mathcal{R}}_{n,\hat{q}}\right) \geq 1 - \alpha,$$

for all random models $\hat{q} \in \mathcal{Q}$ depending on the data.

2. *the simultaneous inference problem is solved, that is,*

$$\mathbb{P}\left(\bigcap_{q \in \mathcal{Q}} \left\{\theta_q \in \hat{\mathcal{R}}_{n,q}\right\}\right) \geq 1 - \alpha.$$

- Only (2) $\Rightarrow$ (1) was proved in Berk et al. (2013).

Solving Simultaneous Inference Problem

To solve the PoSI problem, we introduce the following assumptions:

(A1) $|\mathcal{Q}| \neq \infty$ and

$$\sqrt{\log(e|\mathcal{Q}|)} \max_{q \in \mathcal{Q}} |R_{n,q}| = o_p(1).$$

(A2) The influence functions $\{\psi_{n,q}(\cdot) : q \in \mathcal{Q}\}$ are sub-exponential.

(A3) There exists estimators $\hat{\sigma}^2_{n,q}$ of $\sigma^2_{n,q} = n\text{Var}(\hat{\theta}_q - \theta_q)$ such that

$$\log(e|\mathcal{Q}|) \max_{q \in \mathcal{Q}} \left| \frac{\hat{\sigma}_{n,q}}{\sigma_{n,q}} - 1 \right| = o_p(1).$$

(A4) There exists estimators $\{\hat{\psi}_{n,q} : q \in \mathcal{Q}\}$ of the influence functions satisfying

$$\log(e|\mathcal{Q}|) \max_{q \in \mathcal{Q}} \frac{1}{n} \sum_{i=1}^{n} \left(\hat{\psi}_{n,q}(Z_i) - \psi_{n,q}(Z_i) \right)^2 = o_p(1).$$

Some Comments

- The assumptions allow for a Gaussian approximation result for $\{\sqrt{n}(\hat{\theta}_q - \theta_q), q \in \mathcal{Q}\}$.
- If $|\mathcal{Q}| = 1$, then the assumptions reduce to the classical ones leading to asymptotic normality based inference.
- The assumption of sub-exponential influence functions can be weakened substantially without much difficulty.
- The framework is very closely related to the "*Many Approximate Means* (MAM)" framework of Belloni et al. (2018).
- Although the results can be extended to infinitely many models, we restrict to $|\mathcal{Q}| \neq \infty$ but it can grow with n (almost exponentially).

Some Notation

- Define the concatenated scaled influence function vector as

$$\varphi_{n,\mathcal{Q}}(Z_i) := \left(\frac{\psi_{n,q}(Z_i)}{\sigma_{n,q}} : q \in \mathcal{Q}\right)^\top.$$

- Based on the estimators of influence functions, define

$$\hat{\varphi}_{n,\mathcal{Q}}(Z_i) := \left(\frac{\hat{\psi}_{n,q}(Z_i)}{\hat{\sigma}_{n,q}} : q \in \mathcal{Q}\right)^\top.$$

- Finally, define a Gaussian "process" $G_{n,\mathcal{Q}} \in \mathbb{R}^{|\mathcal{Q}|}$ with mean zero and the covariance given by

$$\mathrm{Var}\left(G_{n,\mathcal{Q}}\right) = \mathrm{Var}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi_{n,\mathcal{Q}}(Z_i)\right).$$

Gaussian Approximation and Bootstrap

Let $\mathcal{A}^{\mathrm{re}}$ be the set of all rectangles in $\mathbb{R}^{|\mathcal{Q}|}$ and $\mathcal{D}_n = \{Z_1, \ldots, Z_n\}$.

Theorem (Belloni et al. (2018) and Kuchibhotla et al. (2018+))

- *Under assumptions (A1)–(A3), if* $\log^7(|\mathcal{Q}|) = o(n)$*, then*

$$\sup_{A \in \mathcal{A}^{re}} \left| \mathbb{P}\left(\left\{ \frac{\sqrt{n}(\hat{\theta}_q - \theta_q)}{\hat{\sigma}_{n,q}} \right\}_{q \in \mathcal{Q}} \in A \right) - \mathbb{P}\left(G_{n,\mathcal{Q}} \in A\right) \right| = o(1).$$

- *Under assumptions (A1)–(A4), if* $\log^7(|\mathcal{Q}|) = o(n)$ *and* $Z_1, \ldots, Z_n$ *are iid, then*

$$\sup_{A \in \mathcal{A}^{re}} \left| \mathbb{P}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} e_i \hat{\varphi}_{n,\mathcal{Q}}(Z_i) \in A \middle| \mathcal{D}_n \right) - \mathbb{P}\left(G_{n,\mathcal{Q}} \in A\right) \right| = o(1).$$

Here $e_1, \ldots, e_n \sim N(0, 1)$.

Sketch of the Proof

- By Assumptions (A1) and (A3),

$$\frac{\sqrt{n}(\hat{\theta}_q - \theta_q)}{\hat{\sigma}_{n,q}} \overset{(A3)}{\approx} \frac{\sqrt{n}(\hat{\theta}_q - \theta_q)}{\sigma_{n,q}} \overset{(A1)}{\approx} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\psi_{n,q}(Z_i)}{\sigma_{n,q}}.$$

The approximations above are uniform in $q \in \mathcal{Q}$. CLT for averages.

- For bootstrap, note

$$\frac{1}{n} \sum_{i=1}^{n} e_i \hat{\varphi}_{n,\mathcal{Q}}(Z_i) \Big| \mathcal{D}_n \sim N\left(0, \frac{1}{n} \sum_{i=1}^{n} \hat{\varphi}_{n,\mathcal{Q}}(Z_i) \hat{\varphi}_{n,\mathcal{Q}}^{\top}(Z_i)\right),$$

and by assumptions (A2) and (A4)

$$\frac{1}{n} \sum_{i=1}^{n} \hat{\varphi}_{n,\mathcal{Q}}(Z_i) \hat{\varphi}_{n,\mathcal{Q}}^{\top}(Z_i) \overset{(A4)}{\approx} \underbrace{\frac{1}{n} \sum_{i=1}^{n} \varphi_{n,\mathcal{Q}}(Z_i) \varphi_{n,\mathcal{Q}}^{\top}(Z_i) \overset{(A2)}{\approx} \mathrm{Var}\left(G_{n,\mathcal{Q}}\right)}_{\text{requires } \mathbb{E}[\phi_{n,q}(Z_i)] = 0 \text{ which uses iid.}}.$$

Outline

An Illustration for Linear Regression: Recap

- Suppose $Z_1 = (X_1, Y_1), \ldots, Z_n = (X_n, Y_n)$ denote independent random vectors in $\mathbb{R}^p \times \mathbb{R}$.
- Recall $\mathcal{M}(k)$ as all submodels of size bounded by k and set

$$\mathcal{Q} := \{(j, M) : j \in M, M \in \mathcal{M}(k)\}.$$

- The collection of functionals is

$$\left\{\beta_{n,M}(j) : q = (j, M) \in \mathcal{Q}\right\}.$$

- Hence, the number of functionals is

$$|\mathcal{Q}| = \sum_{\ell=1}^{k} \ell \binom{p}{\ell} \le \left(\frac{2ep}{k}\right)^k \quad \text{and} \quad |\mathcal{Q}| \ge \left(\frac{p}{k}\right)^k,$$

and so, $|\mathcal{Q}| \asymp (ep/k)^k$ and $\log(|\mathcal{Q}|) \asymp k \log(ep/k)$.

Verification of Assumptions (A1) and (A4)

- For linear regression and smooth M-estimators, assumption (A1) was verified in Kuchibhotla et al. (2018b). The result for OLS is:

$$\max_{|M|\le k}\left\|\sqrt{n}\left(\hat{\beta}_{n,M}-\beta_{n,M}\right)-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{n,M}(Z_i)\right\|_2 = O_p\left(\frac{k\log(ep/k)}{\sqrt{n}}\right),$$

where

$$\psi_{n,M}(Z_i) := (\Sigma_n(M))^{-1}X_i(M)(Y_i - X_i^\top(M)\beta_M),$$

$$\Sigma_n(M) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i(M)X_i^\top(M)].$$

- A trivial estimator of influence function $\psi_{n,M}$ satisfying (A4) is

$$\hat{\psi}_{n,M}(Z_i) := \left(\hat{\Sigma}_n(M)\right)^{-1}X_i(M)(Y_i - \hat{\beta}_M^\top X_i(M)),$$

with $\hat{\Sigma}_n(M)$ representing the Gram matrix for model M.

Verification of (A3)

- Under the assumption of independence without a well-specified model, the estimator for the variance of $\sqrt{n}(\hat{\beta}_M - \beta_M)$ is given by the sandwich:

$$\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\Sigma}_n(M)\right)^{-1} X_i(M)\left(Y_i - \hat{\beta}_M^\top X_i(M)\right)^2 X_i^\top(M)\left(\hat{\Sigma}_n(M)\right)^{-1}.$$

- It can be proved that the sandwich estimator is uniformly close to the true asymptotic variance at the rate:

$$\sqrt{\frac{k\log(ep/k)}{n}} + \frac{(k\log(ep/k))^2}{n}.$$

- Thus the result applies if $k\log(ep/k) = o(n^{1/7})$. Similar result holds for smooth M-estimators based on results of Kuchibhotla et al. (2018b).

Outline

Construction of PoSI Confidence Regions

- The Gaussian approximation result allows for PoSI confidence regions using quantiles of max Gaussians.
- If there is no structure in $\mathcal{Q}$, then a "conventional" construction can be based on the quantiles of

$$\max_{q \in \mathcal{Q}} \left| \frac{\sqrt{n}(\hat{\theta}_q - \theta_q)}{\hat{\sigma}_{n,q}} \right| .$$

- This is called the "max-$|t|$" statistic and was used for PoSI in Berk et al. (2013) and Bachoc et al. (2016).
- BUT, in regression analysis, there is a hierarchical structure:

  model M and then covariate j in model M.

- Ignoring this structure leads to certain deficiencies of the "max-$|t|$" confidence regions.

Deficiencies of max-|t| Regions: Part I

- Define for any model M,

$$T_M := \max_{1\leq j\leq |M|} \left| \frac{\sqrt{n}\left(\hat{\beta}_{n,M}(j) - \beta_{n,M}(j)\right)}{\hat{\sigma}_{n,M}(j)} \right|.$$

- In this notation, max-$|t| := \max_{|M|\leq k} |T_M|$.
- Suppose $M \subset M'$ are two models. Then T_M is usually smaller than $T_{M'}$. Under certain assumptions,

$$\mathbb{E}\left[T_M\right] \asymp \sqrt{\log|M|} \quad \text{and} \quad \mathbb{E}\left[T_{M'}\right] \asymp \sqrt{\log|M'|}.$$

- So, the maximum in the max-|t| is usually *attained at the largest model* implying larger confidence regions for smaller models.

Smaller Models should have Smaller Confidence Regions

Deficiencies of max-|t| Regions: Part II

- To understand the second major deficiency of max-|t| regions, consider the gram matrix

$$\hat{\Sigma}_n := \begin{bmatrix} I_{p-1} & c\mathbf{1}_{p-1} \\ c\mathbf{1}_{p-1}^\top & 1 \end{bmatrix},$$

with $c^2 < 1/(p-1)$ and where $\mathbf{1}_{p-1} = (1, 1, \ldots, 1)^\top$.

- In this setting for most submodels, the covariates are uncorrelated but the full model is highly collinear for $c^2 \approx 1/(p-1)$.
- It was shown in Berk et al. (2013) that max-|t| $\asymp \sqrt{p}$. But if we ignore the last covariate, then max-|t| $\asymp \sqrt{\log p}$.

Collinearity in a model should not affect confidence regions for another model.

Three Confidence Regions

- Recall the max-|t| for model M and standardized max-|t| as

$$T_M := \max_{1\le j\le |M|} \left| \frac{\sqrt{n}\left(\hat{\beta}_M(j) - \beta_M(j)\right)}{\hat{\sigma}_{n,M}(j)} \right|, \text{ and } T_M^\star := \frac{T_M - \mathbb{E}[T_M]}{\sqrt{\mathrm{Var}(T_M)}}.$$

- Consider the following three max statistics:

$$T_k^{(1)} := \max_{|M|\le k} T_M,$$

$$T_k^{(2)} := \max_{1\le s\le k} \left(\frac{\max_{|M|=s} T_M^\star - E_s}{SD_s} \right), \text{ where } E_s := \mathbb{E}\left[\max_{|M|=s} T_M^\star \right],$$

$$T_k^{(3)} := \max_{1\le s\le k} \left(\frac{\max_{|M|\le s} T_M^\star - E_s^\star}{SD_s^\star} \right), \text{ where } E_s^\star := \mathbb{E}\left[\max_{|M|\le s} T_M^\star \right].$$

The quantities SD_s and $SD_s^\star$ are defined similarly to E_s and $E_s^\star$.

Three Confidence Regions Contd.

- The confidence regions (rectangles) are given by

$$\hat{\mathcal{R}}_{n,M}^{(1)} := \left\{\theta : \ T_M(\theta) \le K_\alpha^{(1)}\right\}, \ T_M(\theta) := \max_{j\in M}\left|\frac{\sqrt{n}(\hat{\beta}_M(j) - \theta(j))}{\hat{\sigma}_{n,M}(j)}\right|,$$

$$\hat{\mathcal{R}}_{n,M}^{(2)} := \left\{\theta : T_M(\theta) \le \mathbb{E}[T_M] + \sqrt{\mathrm{Var}(T_M)}(E_s + SD_s K_\alpha^{(2)})\right\},$$

$$\hat{\mathcal{R}}_{n,M}^{(3)} := \left\{\theta : T_M(\theta) \le \mathbb{E}[T_M] + \sqrt{\mathrm{Var}(T_M)}(E_s^\star + SD_s^\star K_\alpha^{(3)})\right\}.$$

  Here $K_\alpha^{(j)}$ denote the quantiles of $T_k^{(j)}$ respectively for $j = 1, 2, 3$.

- The regions $\hat{\mathcal{R}}_{n,M}^{(j)}, j = 2, 3$ provide model dependent scaling and so give **shorter confidence regions for smaller models.**
- Note that all these regions are tight: there exists a model-selection procedure for which the confidence regions have (asymptotically) exact coverage of $1 - \alpha$.

Some Comments

- The three confidence regions provide asymptotically valid post-selection inference.
- Because of the model-dependent scaling for the last two, they are less conservative than the max-|t| confidence regions.
- In most applications with smaller chosen models, the last two confidence regions turn out to be much smaller than the max-|t| confidence regions.
- Bootstrapping $T_k^{(j)}, j = 2, 3$ requires estimation of first two moments of maximums and the results of Banerjee et al. (2018) imply that consistent estimation of moments is possible by Gaussian approximation.
- The three maximum-statistics listed here are not the only options and one can get very creative in designing others.

Outline

Boston Housing Data

Boston housing dataset contains data on $n = 506$ median value of a house along with 13 predictors like crime rate, nitric oxide concentration, number of rooms, percent low status population.

The confidence regions for model $M \in \mathcal{M}(k)$ are given by

$$|T_M(\theta)| \le \begin{cases} K_\alpha^{(1)}, \\ C_M^{(2)} := \mathbb{E}[T_M] + \sqrt{\mathrm{Var}(T_M)}(E_s + SD_s K_\alpha^{(2)}), \\ C_M^{(3)} := \mathbb{E}[T_M] + \sqrt{\mathrm{Var}(T_M)}(E_s^\star + SD_s^\star K_\alpha^{(3)}). \end{cases}$$

We estimate the right hand side quantities using multiplier bootstrap.

To understand how small/wide the last two confidence regions are, we compute:

$$\text{Summary}\left(\frac{C_M^{(2)}}{K_\alpha^{(1)}} : M \in \mathcal{M}(k)\right) \text{ and } \text{Summary}\left(\frac{C_M^{(3)}}{K_\alpha^{(1)}} : M \in \mathcal{M}(k)\right).$$

Boston Housing Data Contd.

Recall that Boston housing data has 14 predictors including the intercept. $1 \leq k \leq 14$ represents the maximum model size allowed and $j = 2, 3$ represents two confidence regions. Here we consider two cases $k = 6$ and $k = 14$.

Table: Comparison of Constants in $\hat{\mathcal{R}}_{n,M}^{(2)}$ and $\hat{\mathcal{R}}_{n,M}^{(3)}$ to max-|t| constant.

Quantiles→		Min.	5%	25%	50%	Mean	75%	95%	Max.
$k = 6$	$j = 2$	0.702	0.978	1.037	1.060	1.052	1.077	1.098	1.140
	$j = 3$	0.692	0.980	1.047	1.072	1.062	1.090	1.112	1.155
$k = 14$	$j = 2$	0.718	0.996	1.044	1.065	1.060	1.083	1.105	1.148
	$j = 3$	0.678	0.999	1.050	1.070	1.064	1.086	1.108	1.147

About 30% gain with about 15% loss over all models!

For $\geq$ 90% of models, the confidence regions are wider.

Outline

Conclusions

- We have provided a unified framework for post-selection inference allowing for increasing number of models.
- We have verified the assumption for OLS and Smooth M-estimators including GLM's.
- Based on the Gaussian approximation results, we have constructed and implemented three different PoSI confidence regions.
- All three confidence regions are asymptotically tight. This implies that no one can uniformly dominate the other.
- An interesting question of what kind of maximum statistic to consider is raised.
- Efficienct algorithms and detailed simulation studies are under progress.

References

[1] Bachoc, F., Preinerstorfer, D., and Steinberger, L. (2016).
Uniformly valid confidence intervals post-model-selection.
arxiv.org/abs/1611.01043.

[2] Banerjee, D., Kuchibhotla, A. K., and Mukherjee, S. (2018).
Cramér-type large deviation and non-uniform central limit theorems in high dimensions.
arxiv.org/abs/1806.06153.

[3] Belloni, A., Chernozhukov, V., Chetverikov, D., Hansen, C., Kato, K., et al. (2018).
High-dimensional econometrics and regularized gmm.
Technical report, Centre for Microdata Methods and Practice, Institute for Fiscal Studies.

[4] Berk, R., Brown, L. D., Buja, A., Zhang, K., Zhao, L. (2013)
Valid post-selection inference.
Ann. Statist. 41, no. 2, 802–837.

[5] Kuchibhotla, Brown, Buja, Berk, Zhao, George (2018a)
Valid Post-selection Inference in Assumption-lean Linear Regression.
arxiv.org/abs/1806.04119.

[6] Kuchibhotla, Brown, Buja, Zhao, George (2018b)
A Model Free Perspective for Linear Regression: Uniform-in-model Bounds for Post Selection Inference.
arxiv.org/abs/1802.05801.

[7] Kuchibhotla, Brown, Buja, Zhao, George (2018+)
A Note on Post-selection Inference for M-estimators.
in preparation.

Thank You
Questions?

Notation for a General Result

- Suppose (X_i, Y_i) be n random vectors and for $M \subseteq \{1, 2, \ldots, p\}$, define the estimator

$$\hat{\beta}_{n,M} := \underset{\theta \in \mathbb{R}^{|M|}}{\arg\min} \; \frac{1}{n} \sum_{i=1}^{n} L\left(Y_i, X_i^\top(M)\theta\right),$$

and the corresponding target

$$\beta_{n,M} := \underset{\theta \in \mathbb{R}^{|M|}}{\arg\min} \; \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[L\left(Y_i, X_i^\top(M)\theta\right)\right].$$

Here $L(\cdot, \cdot)$ is a loss function convex in the second argument.

- GLM's are special case with $L(y, t) = L(y, t) = \psi(t) - yt$; logistic: $\psi(u) = \log(1 + \exp(u))$ and Poisson: $\psi(u) = \exp(u)$

Notation Contd.

- Set

$$L'(y,u) := \frac{\partial}{\partial t} L(y,t)\Big|_{t=u} \quad \text{and} \quad L''(y,u) := \frac{\partial}{\partial t} L'(y,t)\Big|_{t=u}.$$

- Define

$$C_{+}(y,u) := \sup_{|s-t|\le u} \frac{L''(y,s)}{L''(y,t)} \quad (\ge 1).$$

For logistic and Poisson, $C_{+}(y,u) \le \exp(3u)$ for all y.

- Finally, define the estimating function for model M as

$$\hat{\mathcal{Z}}_{n,M}(\theta) := \frac{1}{n}\sum_{i=1}^{n} L'\left(Y_i, X_i^{\top}(M)\theta\right) X_i(M) \in \mathbb{R}^{|M|}$$

$$\hat{\mathcal{J}}_{n,M}(\theta) := \frac{1}{n}\sum_{i=1}^{n} L''\left(Y_i, X_i^{\top}(M)\theta\right) X_i(M) X_i^{\top}(M) \in \mathbb{R}^{|M|\times|M|}.$$

A Main Result: Deterministic Version

Theorem

For any $n, k \geq 1$ and for $M \subseteq \{1, 2, \ldots, p\}$, set

$$\delta_{n,M} := \left\| \left[\hat{\mathcal{J}}_{n,M}(\beta_{n,M}) \right]^{-1} \hat{\mathcal{Z}}_{n,M}(\beta_{n,M}) \right\|_2 ,$$

and the event

$$\mathcal{E}_{k,n} := \left\{ \max_{|M| \leq k} \max_{1 \leq i \leq n} C_+(Y_i, 2\|X_i(M)\|_2 \, \delta_{n,M}) \leq \frac{3}{2} \right\}.$$

On the event $\mathcal{E}_{k,n}$, simultaneously for all models $|M| \leq k$, there exists a unique $\hat{\beta}_{n,M} \in \mathbb{R}^{|M|}$ satisfying

$$\hat{\mathcal{Z}}_{n,M}(\hat{\beta}_{n,M}) = 0 \quad \text{and} \quad \frac{1}{2}\delta_{n,M} \leq \left\| \hat{\beta}_{n,M} - \beta_{n,M} \right\|_2 \leq 2\delta_{n,M}.$$

Uniform Linear Representation

Theorem

On the event $\mathcal{E}_{k,n}$, simultaneously for all models $|M| \leq k$, the estimators satisfy

$$\left\| \hat{\beta}_{n,M} - \beta_{n,M} + \left[\mathcal{J}_{n,M}(\beta_{n,M})\right]^{-1} \hat{\mathcal{Z}}_{n,M}(\beta_{n,M}) \right\|_2 \leq \Delta_{n,M}\delta_{n,M},$$

where

$$\mathcal{J}_{n,M}(\theta) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[L''\left(Y_i, X_i^\top(M)\theta\right) X_i(M)X_i^\top(M)\right],$$

and

$$\Delta_{n,M} = \frac{\left\| \hat{\mathcal{J}}_{n,M}(\beta_{n,M}) - \mathcal{J}_{n,M}(\beta_{n,M}) \right\|_{op}}{\lambda_{\min}\left(\mathcal{J}_{n,M}(\beta_{n,M})\right)} + \max_i C_+(Y_i, 2\left\|X_i(M)\right\|_2 \delta_{n,M}) - 1.$$

Note that independence of observations is NOT required.

Application

- For generalized linear models (in the canonical form),

$$L(y, t) = \psi(t) - yt,$$

for a convex function $\psi(\cdot)$.

- In logistic and Poisson regression, the event $\mathcal{E}_{k,n}$ becomes

$$\max_{|M|\le k} \max_{1\le i\le n} \|X_i(M)\|_2 \, \delta_{n,M} \le \frac{\log 2}{6}. \tag{1}$$

- No independence is required.
- Inequality (1) holds as long as $k = o(\sqrt{n/\log p})$ under the tail assumption and "weak" dependence.
- The proof is based on the Banach fixed point theorem and also applies to Cox proportional hazards model.