Construction of PoSI Statistics[1]

Andreas Buja and Arun Kumar Kuchibhotla

Department of Statistics
University of Pennsylvania

September 8, 2018

WHOA-PSI 2018

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

In Memory of

LAWRENCE D. BROWN †

1940 – 2018.

Outline

Outline

A Crisis in the Sciences: Irreproducibility

- Indicators of a crisis:
  - Bayer Healthcare reviewed 67 in-house attempts at replicating findings in published research: $< 1/4$ were viewed as replicated
  - Arrowsmith (2011, Nat. Rev. Drug Discovery 10):
    Increasing failure rate in Phase II drug trials
  - Ioannidis (2005, PLOS Medicine):
    "Why Most Published Research Findings Are False"
  - Simmons, Nelson, Simonsohn (2011, Psychol.Sci):
    "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant"

$\Longrightarrow$ **Irreproducibility of Empirical Findings**

- Many potential causes – two major ones:
  - Institutional: Publication bias, "file drawer problem"
  - Methodological: Statistical biases, **"researcher degrees of freedom"**

Irreproducibility: Methodol. Factor 1 – Selection

- A statistical bias is due to lack of accounting for
  selection of variables, transforms, scales, subsets, weights, ...
- **Regressor/model selection** (our focus) is on several levels:
  - formal selection: all subset (C_p, AIC, BIC,...), stepwise (F), lasso,...
  - informal selection: diagnostics for GoF, influence, collinearity,...
  - post hoc selection: "Effect size is too small, the variable too costly."
- Suspicions and Criticisms:
  - All three modes of selection are (should be) used.
  - More thorough data analysis $\Longrightarrow$ More spurious results
  - Not a solution: Post-selection inference for "adaptive Lasso", say.
    Empirical researchers do not write contracts with themselves to commit **a priori** to one formal selection method and nothing else.

The “PoSI” Solution to Selection: FWER Control

- PoSI Procedure — general version:
  - Define a **universe** $\mathcal{M}$ of models M you might ever consider/select: outcomes (Y), regressors (X), their transforms ($f(X), g(Y)$), ...
  - Define the universe of all tests you might ever perform in these models, typically for regression coeffs $\beta_{j,M}$ (j'th coeff in model M).
  - Consider the **minimum of the p-values** for all these tests: Obtain its 0.05 quantile $\alpha_{0.05}$ for FWER adjustment.
  - Now freely examine your data and select models $\hat{M} \in \mathcal{M}$, reconsider, re-select, re-reconsider, ... but compare all p-values against $\alpha_{0.05}$, not 0.05, for **0.05 $\mathcal{M}$-FWER control**.
- Cost-Benefit Analysis:
  - Cost: Huge computation upfront — adjustment for millions of tests
  - Benefits: **Solution to the circularity problem** — select model $\hat{M}$, don't like it, select $\hat{M}'$, don't like it, ... PoSI inference remains valid.

- Models are approximations, not generative truths.
  $\Longrightarrow$ Consequences!
- What is the target $\beta_{j,M}$ of $\hat{\beta}_{j,M}$? Stay tuned.
- Model bias interacts with regressor distributions to cause model-trusting SEs to be off, sometimes too small by a factor of 2.

$$V[\hat{\beta}] \;=\; E[V[\hat{\beta}|X]] + V[E[\hat{\beta}|X]]$$

- Do not condition on the regressors; do not treat them as fixed!
- Use model-robust standard errors, for example, from the x-y pairs or multiplier bootstraps, **not** the residual bootstrap!

Wanted: PoSI Protection under Misspecification!

Up next: PoSI under Misspecification and PoSI Statistic

Outline

For models $M \subseteq \{1, 2, \ldots, p\}$ and IID random vectors (X_i, Y_i). For any function f, set

$$\hat{\mathbb{P}}_n[f(X, Y)] = \frac{1}{n}\sum_{i=1}^{n} f(X_i, Y_i), \quad \text{and} \quad \mathbb{P}[f(X, Y)] = \mathbb{E}[f(X_1, Y_1)].$$

Sample

- Gram matrix:

$$\hat{\Sigma}_n := \hat{\mathbb{P}}_n\left[XX^\top\right].$$

- "Covariance" Vector:

$$\hat{\Gamma}_n := \hat{\mathbb{P}}_n[XY].$$

- Estimator:

$$\boxed{\hat{\beta}_{n,M} := (\hat{\Sigma}_n(M))^{-1}\hat{\Gamma}_n(M).}$$

Population

- Gram matrix:

$$\Sigma := \mathbb{P}\left[XX^\top\right].$$

- "Covariance" Vector:

$$\Gamma := \mathbb{P}[XY].$$

- Target:

$$\boxed{\beta_M := (\Sigma(M))^{-1}\,\Gamma(M).}$$

Uniform-in-submodel Result for OLS

If $Z_i := (X_i, Y_i)$ are *sub-Gaussian*, then the results of Kuchibhotla et al. (2018b) imply that for any $1 \le k \le p$,

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

and

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

where

$$\psi_M(Z_i) := (\Sigma(M))^{-1} X_i(M)(Y_i - X_i^\top(M)\beta_M).$$

Recall

$$\Sigma(M) = \mathbb{E}[X_1(M) X_1^\top(M)] \quad \text{and} \quad \beta_M := (\Sigma(M))^{-1} \mathbb{E}[X_1(M) Y_1].$$

Implications for PoSI

- These results imply that if $k \log(ep/k) = o(\sqrt{n})$, then as $n \to \infty$, simultaneously for all $|M| \le k$,

$$\sqrt{n}\left(\hat{\beta}_{n,M} - \beta_M\right) \;\approx\; \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_M(Z_i).$$

- This implies one can apply bootstrap to estimate quantiles of the "max-|t|" statistic:

$$\text{max-|t|} := \max_{|M|\le k,\, j\in M}\left|\frac{\sqrt{n}(\hat{\beta}_{n,M}(j) - \beta_M(j))}{\hat{\sigma}_M(j)}\right|.$$

- The linear representation result holds also for functionally dependent and/or non-identically distributed observations. See Kuchibhotla et al. (2018b).

Outline

Is max-|t| the right statistic??

- The "max-$|t|$" statistic was used for PoSI in Berk et al. (2013) and Bachoc et al. (2016).
- max-|t| is definitely not the only choice. So, can we do any better?
- 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.

To follow: Deficiencies of max-|t| and new PoSI Regions

Deficiencies of max-|t| Regions: Part I

- Define

$$T_M := \max_{j \in M} \left| \frac{\sqrt{n}\left(\hat{\beta}_{n,M}(j) - \beta_M(j)\right)}{\hat{\sigma}_M(j)} \right| \quad \text{and} \quad \text{max-}|t| := \max_{|M| \le k} |T_M|.$$

- 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: Part I contd.

- For any two (fixed) models M, M', as $n \to \infty$,

$$\max\left\{\sqrt{\frac{n}{|M|}}\left\|\hat{\beta}_{n,M} - \beta_M\right\|_2, \sqrt{\frac{n}{|M'|}}\left\|\hat{\beta}_{n,M'} - \beta_{M'}\right\|_2\right\} = O_p(1).$$

- So, without model selection smaller models have smaller confidence regions.
- But the max-|t| confidence regions do NOT maintain this.
- This is of importance especially if the total number of covariates is larger than the sample size.

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.

How to Remedy this: A Simplified Example

- Suppose $W_j \sim N(\mu_j, 1)$ for $1 \leq j \leq p$. We want PoSI for $\mu_{\hat{j}}$ for $\hat{j}$ chosen based on the sequence.
- If W_j are independent, then max-|t| confidence region for $\mu_{\hat{j}}$ is essentially

$$\{\theta : |W_{\hat{j}} - \theta| \leq (2 \log p)^{1/2}\}.$$

- Is this the best?? Consider the statistic

$$S^\star := \max_{1 \leq j \leq p} \frac{|W_j - \theta_j|}{(2 \log(j))^{1/2}} \qquad \leftarrow \text{index dependent scaling.}$$

- It is easy to prove that $S^\star = O_p(1)$ (even if $p = \infty$). This statistic implies the confidence region for $\mu_{\hat{j}}$:

$$\left\{\theta : |W_{\hat{j}} - \theta| \leq C(2 \log(\hat{j}))^{1/2}\right\}$$

- This is much less conservative if the chosen $\hat{j}$ is not too big.

Simplified Example Contd.

- Once again the confidence regions are

$$\{\theta : |W_{\hat{j}} - \theta| \le (2\log p)^{1/2}\}, \tag{1}$$

$$\{\theta : |W_{\hat{j}} - \theta| \le C(2\log(\hat{j}))^{1/2}\}. \tag{2}$$

- Confidence region (2) is uniformly better than (1) (rate-wise).
- Furthermore, both regions (1) and (2) are tight, i.e., there is a $\hat{j}$ such that

$$|W_{\hat{j}} - \mu_{\hat{j}}| = \max_{1\le j\le p} |W_j - \mu_j|,$$

and there is also a $\hat{j}$ such that

$$\frac{|W_{\hat{j}} - \mu_{\hat{j}}|}{(2\log(\hat{j}))^{1/2}} = \max_{1\le j\le p} \frac{|W_j - \mu_j|}{(2\log(j))^{1/2}}.$$

- **Moral of the story:** layer-by-layer standardization helps.

Three Confidence Regions

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

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

- 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.

- Define for $\theta \in \mathbb{R}^{|M|}$,

$$T_M(\theta) := \max_{j \in M} \left| \sqrt{n}(\hat{\beta}_M(j) - \theta(j))/\hat{\sigma}_{n,M}(j) \right|,$$

  and consider the confidence regions (rectangles) are given by

$$\hat{\mathcal{R}}^{(1)}_{n,M} := \left\{ \theta : \ T_M(\theta) \le K^{(1)}_\alpha \right\},$$
$$\hat{\mathcal{R}}^{(2)}_{n,M} := \left\{ \theta : T_M(\theta) \le \mathbb{E}[T_M] + \sqrt{\mathrm{Var}(T_M)}(E_s + SD_s K^{(2)}_\alpha) \right\},$$
$$\hat{\mathcal{R}}^{(3)}_{n,M} := \left\{ \theta : T_M(\theta) \le \mathbb{E}[T_M] + \sqrt{\mathrm{Var}(T_M)}(E^\star_s + SD^\star_s K^{(3)}_\alpha) \right\}.$$

  Here $K^{(j)}_\alpha$ denote the quantiles of $T^{(j)}_k$ respectively for $j = 1, 2, 3$.

- The quantiles $K^{(j)}_\alpha$ can be estimated using multiplier bootstrap (where one replaces $\mathbb{E}[T_M], \mathrm{Var}(T_M), E_s, SD_s$ by their estimators).

Some Comments

- The three confidence regions provide asymptotically valid post-selection inference.
- The regions $\hat{\mathcal{R}}_{n,M}^{(j)}, j = 2, 3$ provide model dependent scaling and so give **shorter confidence regions for smaller models.**
- Because of the model-dependent scaling for the last two, they are less conservative than the max-|t| confidence regions.
- The three maximum-statistics listed here are not the only options and one can get very creative in designing others.

Outline

Boston Housing Data

The Boston housing dataset contains data on $n = 506$ median value of a house along with 13 predictors.
The confidence regions for model $M \in \mathcal{M}(k)$ are given by

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

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

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

This tells for what proportion of models are the second and third regions shorter/wider and by how much?

Boston Housing Data Contd.

There are 14 predictors including the intercept. $k \in \{1, \ldots, 14\}$ represents the maximum model size allowed and $j = 2, 3$ represents the last two confidence regions.

Here we consider two cases $k = 6$ and $k = 14$ (NO file drawer prob.!!).

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 post-selection inference allowing for increasing number of models for linear regression.
- 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.
- A generally interesting question: What kind of maximum statistic should be to consider?
- Efficient algorithms and detailed simulation studies are under development.

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?