Recent Developments in Post-selection Inference (from Larry's Group)[1]

Arun Kumar Kuchibhotla

Department of Statistics
University of Pennsylvania

30 November 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

Some History of PoSI

Some History

- In applied statistics, a formal model is built after a thorough exploration of data.
- Reproducibility/replicability crisis in science is sometimes attributed to this type of data analysis.
- Model Selection/Cherry-picking makes classical statistical inference methods invalid.
- Berk et al. (2013) provided valid statistical inference for Gauss-Markov linear model under arbitrary variable selection.
- However, model misspecifcation also makes classical statistical inference methods invalid.

Some History

- The practice of data analysis often involves exploring the data thoroughly before a formal modeling begins. EDA is an example.
- Reproducibility/replicability crisis in science is sometimes attributed to this type of data analysis.
- **Model Selection/Cherry-picking** makes classical statistical inference methods invalid.
- Berk et al. (2013) provided valid statistical inference for Gauss-Markov linear model under arbitrary variable selection.
- However, model **misspecification** also makes classical statistical inference methods invalid.

Wanted: Valid inference under misspecification and model selection!

The "PoSI" Solution of Berk et al. (2013): Simultaneity

- 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 **maximum of the test statistics** for all these tests: Obtain its 0.05 critical value $\boldsymbol{C}_{0.05}$ for **simultaneity** adjustment.
  - Now freely examine your data and select models $\hat{M} \in \mathcal{M}$, reconsider, re-select, re-reconsider, ... but compare all statistics against $\boldsymbol{C}_{0.05}$, for **0.05 $\mathcal{M}$-simultaneity 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.

Equivalence of PoSI and Simultaneous Inference

- For any set of functionals $\{\theta_M : M \in \mathcal{M}\}$ and confidence regions $\{\hat{\mathcal{R}}_M : M \in \mathcal{M}\}$, it is clear that **for any $\hat{M} \in \mathcal{M}$,**

$$\mathbb{P}\left(\theta_{\hat{M}} \in \hat{\mathcal{R}}_{\hat{M}}\right) \geq \mathbb{P}\left(\bigcap_{M\in\mathcal{M}} \{\theta_M \in \hat{\mathcal{R}}_M\}\right),$$

Post-selection Inf. $\Leftarrow$ **Simultaneous Inf.**

- We have proved (Kuchibhotla et al. (2018a)) that

$$\inf_{\hat{M}\in\mathcal{M}} \mathbb{P}\left(\theta_{\hat{M}} \in \hat{\mathcal{R}}_{\hat{M}}\right) = \mathbb{P}\left(\bigcap_{M\in\mathcal{M}} \{\theta_M \in \hat{\mathcal{R}}_M\}\right),$$

Post-selection Inf. $\Leftrightarrow$ **Simultaneous Inf.**

- Thus, simultaneous inference is necessary and sufficient for PoSI. All our methods aim for simultaneous inference.

Computationally Efficient PoSI for OLS

Notation

- Suppose $(X_i, Y_i) \in \mathbb{R}^p \times \mathbb{R}, 1 \le i \le n$ are observations that constitute regression data.
- Let the data matrices be

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} \in \mathbb{R}^{n \times p} \quad \text{and} \quad \mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \in \mathbb{R}^n.$$

- For any $1 \le k \le p$, let

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

represent the set of ***k*-sparse models**.

Notation Contd.

- The OLS least squares estimator based on $(\mathbf{X}(M), \mathbf{Y})$ is given by

$$\hat{\beta}_M = (\mathbf{X}_M^\top \mathbf{X}_M)^{-1}(\mathbf{X}_M \mathbf{Y}).$$

- The OLS least squares target based on $(\mathbf{X}(M), \mathbf{Y})$ is given by

$$\beta_M = (\mathbb{E}[\mathbf{X}_M^\top \mathbf{X}_M])^{-1}(\mathbb{E}[\mathbf{X}_M \mathbf{Y}])$$

- Let $C_{xx}(\alpha), C_{xy}(\alpha)$ be such that with probability $1 - \alpha$,

$$\left\{ \left\| \frac{\mathbf{X}^\top \mathbf{X} - \mathbb{E}[\mathbf{X}^\top \mathbf{X}]}{n} \right\|_\infty \le C_{xx}(\alpha) \;\; \& \;\; \left\| \frac{\mathbf{X}\mathbf{Y} - \mathbb{E}[\mathbf{X}\mathbf{Y}]}{n} \right\|_\infty \le C_{xy}(\alpha) \right\},$$

holds. Hence, $C_{xx}(\alpha), C_{xy}(\alpha)$ denote the quantiles of the joint distribution.

Computationally Efficient Valid PoSI

- Consider for any $M \subseteq \{1, 2, \ldots, p\}$, the region

$$\hat{\mathcal{R}}_M := \left\{\theta \in \mathbb{R}^{|M|} : \|\hat{\Sigma}_{n,M}(\hat{\beta}_M - \theta)\|_\infty \leq \mathbf{C_{xy}}(\alpha) + \mathbf{C_{xx}}(\alpha)\|\hat{\beta}_M\|_1\right\},$$

  where $\hat{\Sigma}_{n,M} := \mathbf{X}_M^\top \mathbf{X}_M/n$.
- For **independent** or **functionally dependent sub-Gaussian** observations, if $k = o(\sqrt{n/\log p})$, then

$$\liminf_{n\to\infty} \mathbb{P}\left(\bigcap_{M\in\mathcal{M}(k)} \left\{\beta_M \in \hat{\mathcal{R}}_M\right\}\right) \geq 1 - \alpha. \qquad \text{(Valid PoSI)}$$

- These are polyhedral confidence regions parallelepiped in shape.
- Under above conditions, as $n \to \infty$,

$$\max\left\{C_{xx}(\alpha), C_{xy}(\alpha)\right\} = O\left(\sqrt{\frac{\log p}{n}}\right).$$

Computation

- To compute the confidence region $\hat{\mathcal{R}}_{\hat{M}}$ for a model $\hat{M}$, we only need to compute

$$\hat{\Sigma}_{n,\hat{M}},\ \hat{\beta}_{\hat{M}},\ C_{xx}(\alpha), C_{xy}(\alpha).$$

- The first two are readily available for computations leading to $\hat{\beta}_{\hat{M}}$.
- The last two do not depend on $\hat{M}$ and only require computations of order p^2; under independence they are obtained by generating

$$\left\|\left\|\left\| \frac{1}{n}\sum_{i=1}^{n} Z_i (X_i X_i^\top - \hat{\Sigma}_n) \right\|\right\|\right\|_\infty \quad \text{and} \quad \left\| \frac{1}{n}\sum_{i=1}^{n} Z_i (X_i Y_i - \mathbf{Ave}(XY)) \right\|_\infty,$$

where $Z_1, \ldots, Z_n \overset{iid}{\sim} N(0,1)$. This is called **Multiplier Bootstrap**.
- Bootstrap asymptotics require $\log^5 p = o(n)$.
- A similar bootstrap works under functional dependence.

Lebesgue Measures

The PoSI guarantee does not require observations to be identically distributed; so covers the case of fixed design.

Reference	**Leb**$(\hat{\mathcal{R}}_{\hat{M}})$	Design
Kuchibhotla et al. (2018a)	$(\log p/n)^{\lvert\hat{M}\rvert/2}$	fixed design
	$(\lvert\hat{M}\rvert \log p/n)^{\lvert\hat{M}\rvert/2}$	random design
Berk et al. (2013) Bachoc et al. (2016) Kuchibhotla et al. (2018b)	$(k \log p/n)^{\lvert\hat{M}\rvert/2}$	fixed/random design
Taylor and Co. (2016+)	Infinite	fixed/random design

Table: Lebesgue Measures of Different PoSI Regions over models $M \in \mathcal{M}(k)$.

Statistically Tight PoSI for OLS

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 $1 \le k \le p$,

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

and

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

where

$$\psi_M(Z_i) := \Sigma_M^{-1} X_{i,M}(Y_i - X_{i,M}^\top \beta_M).$$

Recall

$$\Sigma_M = \mathbb{E}\left[\frac{\mathbf{X}_M^\top \mathbf{X}_M}{n}\right] \quad \text{and} \quad \beta_M := \Sigma_M^{-1}\mathbb{E}\left[\frac{\mathbf{X}_M \mathbf{Y}}{n}\right].$$

Implications for PoSI

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

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

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

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

  Here $\hat{\sigma}_M(j)$ represents an estimate of the standard error.

- This leads to an asymptotically tight PoSI in that there exists a model selection procedure for which smaller confidence regions are invalid.

Conclusions

Conclusions

- We have provided post-selection inference allowing for increasing number of models for OLS linear regression.
- Based on the Gaussian approximation results, we have constructed and implemented two different PoSI regions.
- The first set of regions are computationally efficient: $\log p = o(n^{\frac{1}{5}})$.
- The second set of regions are statistically tight: $k \log(\frac{ep}{k}) = o(n^{\frac{1}{5}})$.
- Approximate (heuristic) methods for statistically tight regions are under study.
- Similar results holds for a large class of M-estimators and the methodology readily allows for explorations other than variable selection like transformations.

References

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

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

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

[4] 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.

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

Thank You and
Thanks to Linda Zhao and Junhui (Jeff) Cai.