

Uniformly valid confidence intervals post-model-selection

François Bachoc, David Preinerstorfer and Lukas Steinberger

Institut de Mathématiques de Toulouse
Université Libre de Bruxelles
University of Freiburg

Toulouse - Potsdam
Stochastic processes and statistical machine learning, II
March 2019

- 1 The post-model selection inference setting
- 2 The confidence intervals for Gaussian homoscedastic linear models
- 3 The confidence intervals for more general situations
- 4 Some simulation results

Data :

- We consider a triangular array of independent $1 \times l$ random vectors $y_{1,n}, \dots, y_{n,n}$
- We let $\mathbb{P}_n = \bigotimes_{i=1}^n \mathbb{P}_{i,n}$ be the distribution of $y_n = (y'_{1,n}, \dots, y'_{n,n})'$, where $\mathbb{P}_{i,n}$ is the distribution of $y_{i,n}$

Models :

- We now consider a set $M_n = \{\mathbb{M}_{1,n}, \dots, \mathbb{M}_{d,n}\}$ composed of d **models**
- $\mathbb{M}_{i,n}$ is a set of distributions on $\mathbb{R}^{n \times \ell}$
- d does not depend on n (fixed-dimensional asymptotics)

\implies We do not assume that the observation distribution \mathbb{P}_n belongs to one of the $\{\mathbb{M}_{1,n}, \dots, \mathbb{M}_{d,n}\}$. The set of models can be **misspecified**

Parameters :

- We define for each model $\mathbb{M} \in \mathcal{M}_n$ an **optimal parameter** $\theta_{\mathbb{M},n}^* = \theta_{\mathbb{M},n}^*(\mathbb{P}_n)$, that we assume to be non-random and of fixed dimension $m(\mathbb{M})$
- Typically, $\mathbb{M} \in \mathcal{M}_n$ is a set of distributions parameterized by $\theta_{\mathbb{M}} \in \mathbb{R}^{m(\mathbb{M})}$, and $\theta_{\mathbb{M},n}^*$ corresponds to the projection of \mathbb{P}_n on \mathbb{M} , for some distance
- The optimal parameter $\theta_{\mathbb{M},n}^*$ is specific to the model \mathbb{M}

Estimators :

- We consider, for each $\mathbb{M} \in \mathcal{M}_n$, an estimator $\hat{\theta}_{\mathbb{M},n}$ of the optimal parameter $\theta_{\mathbb{M},n}^*$

Example : binary regression 1

Data :

- $l = 1$: scalar observations
- $n \times 1$ observation vector

$$y_n = \begin{pmatrix} y_{1,n} \\ \vdots \\ y_{n,n} \end{pmatrix}$$

- independent components
- $y_i \in \{0, 1\}$
- For $i = 1, \dots, n$, $\mathbb{P}(y_{i,n} = 1) \in [\delta, 1 - \delta]$ for fixed $\delta > 0$ (technical for asymptotics)

$\implies \mathbb{P}_n$ is a distribution on $\{0, 1\}^n$ with independent components and non-vanishing 'randomness'

Example : binary regression 2

Models :

- Let X_n be a $n \times p$ design matrix
- Let $X_{j,n}$ be the j th row of X_n
- Each model \mathbb{M}_j is identified by
 - a set of variables $M_j \subset \{1, \dots, p\}$
 - a response function $h_j : \mathbb{R} \rightarrow [0, 1]$
- Under model \mathbb{M}_j we assume that for $j = 1, \dots, n$

$$\mathbb{P}(y_{j,n} = 1) = h_j(X_{j,n}[M_j]\theta_{\mathbb{M}_j})$$

- for some $|M_j| \times 1$ vector $\theta_{\mathbb{M}_j}$
- with $X_{j,n}[M_j]$ the j th line of $X_n[M_j]$
- where $X_n[M_j]$ is obtained by keeping the columns of X_n with indices in M_j
- we also assume independent components

$\implies \mathbb{M}_j$ is the set of distributions on \mathbb{R}^n with independent components in $\{0, 1\}$ and with mean vector in $h_j(\text{span}(X_n[M_j]))$

Target :

- For a model \mathbb{M}

$$\theta_{\mathbb{M},n}^* \in \operatorname{argmin}_{\theta_{\mathbb{M}} \in \mathbb{R}^{|\mathbb{M}|}} \operatorname{KL}(\mathbb{P}_{\mathbb{M},\theta_{\mathbb{M}}}, \mathbb{P}_n),$$

with

- $\mathbb{P}_{\mathbb{M},\theta_{\mathbb{M}}}$ the distribution in model \mathbb{M} with parameter $\theta_{\mathbb{M}}$
- \mathbb{P}_n the true distribution of the observation vector

Estimator :

- $\hat{\theta}_{\mathbb{M},n}$: the maximum likelihood estimator in the model \mathbb{M}

Model selection :

- We consider a **model selection procedure** : a function $\hat{M}_n : \mathbb{R}^{n \times \ell} \rightarrow M_n$
- We are hence interested in constructing confidence intervals for the random quantity of interest $\theta_{\hat{M}_n, n}^*$
- This is the **post-model-selection** inference framework

Related work :

- [Van der Geer et al. 2014](#), **AoS** lasso for linear models
- [Lee et al. 2016](#), **AoS** lasso for linear models
- [Taylor and Tibschirani 2017](#), **CJoS** lasso for generalized linear models
- [Berk et al 2013](#), **AoS** any model selector for Gaussian linear models

- 1 The post-model selection inference setting
- 2 The confidence intervals for Gaussian homoscedastic linear models**
- 3 The confidence intervals for more general situations
- 4 Some simulation results

- Gaussian vector of observations

$$y_n = \begin{pmatrix} y_{1,n} \\ \vdots \\ y_{n,n} \end{pmatrix} = \mu + \sigma^2 U$$

with

μ fixed and unknown and $U \sim \mathcal{N}(0, I_n)$

- Homoscedastic linear models, with Gaussian errors
 - Observed $n \times p$ design matrix X_n
 - For $M \subset \{1, \dots, p\}$, $|M| \leq n$, $X_n[M]$ corresponds to selecting columns in M
 - Model \mathbb{M} , defined by M , with $|M| \times 1$ parameter $\theta_{\mathbb{M}}$ assumes that

$$y_n = X_n[M]\theta_{\mathbb{M}} + \mathcal{N}(0, \sigma^2 I_n)$$

- We consider that $M \in \mathcal{I} \subset \{M; M \subset \{1, \dots, p\}\}$

- $\theta_{M,n}^*$ corresponds to the projection of the true mean vector on $\text{span}(X_n[M])$ with $M \subset \{1, \dots, p\}$

$$\theta_{M,n}^* = (X_n[M]' X_n[M])^{-1} X_n[M]' \mu$$

- $\theta_{M,n}^*$ is the model-dependent target of inference
- $\hat{\theta}_{M,n}$ is the least square estimator based on $\text{span}(X_n[M])$

$$\hat{\theta}_{M,n} = (X_n[M]' X_n[M])^{-1} X_n[M]' y_n$$

Confidence intervals :

- Berk et al 2013, AoS observe that $\{\hat{\theta}_{\mathbb{M},n} - \theta_{\mathbb{M},n}^*\}_{\mathbb{M} \in \mathcal{I}}$ is Gaussian
- They use a **worst case** approach (in terms of the selected model) and obtain a family of confidence intervals

$$\left\{ \text{CI}_{1-\alpha, \mathbb{M}}^{(j)} \right\}_{\mathbb{M} \in \mathbb{M}_n, j=1, \dots, m(\mathbb{M})},$$

satisfying

$$\mathbb{P}_n \left(\left[\theta_{\hat{\mathbb{M}}_n, n}^* \right]_j \in \text{CI}_{1-\alpha, \hat{\mathbb{M}}_n}^{(j)} \text{ for all } j = 1, \dots, m(\hat{\mathbb{M}}_n) \right) \geq 1 - \alpha$$

Universality :

- holds for any model selector $\hat{\mathbb{M}}_n$: universally valid (Berk et al.)
- particularly beneficial when the statistician has **limited control** on the model selection procedure : informal , cost-driven...

Case $n < p$ (for concision)

- The confidence intervals of [Berk et al 2013](#) are based on computing quantiles of

$$\max_{\substack{M \in \mathcal{I} \\ i=1, \dots, |M|}} \frac{\text{line } i \text{ of } (X_n[M]' X_n[M])^{-1} X_n[M]'}{\text{norm of line } i \text{ of } (X_n[M]' X_n[M])^{-1} X_n[M]'} U,$$

where $U \sim \mathcal{N}(0, I_n)$

- Typically $\mathcal{I} = \{M \subset \{1, \dots, p\}; |M| \leq s\}$ with **sparsity** s

Several challenges :

- Sampling the maximum of a **high-dimensional** Gaussian vector with **small rank** covariance matrix
- Asymptotic behavior of the quantiles ? As a function of X_n ? Some results in [Berk et al. 2013](#), [Bachoc Leeb Pötscher 2019](#), [Bachoc Neuvial Blanchard 2019](#)

- 1 The post-model selection inference setting
- 2 The confidence intervals for Gaussian homoscedastic linear models
- 3 The confidence intervals for more general situations**
- 4 Some simulation results

Main idea :

- We aim at showing a **joint asymptotic normality** of $\{\hat{\theta}_{\mathbb{M},n} - \theta_{\mathbb{M},n}^*\}_{\mathbb{M} \in \mathbb{M}_n}$
- We then use the same construction as in [Berk et al](#) for the confidence intervals
- Additional difficulty : we do not know the asymptotic covariance matrix

Notation :

- $\hat{\theta}_n = (\hat{\theta}'_{\mathbb{M}_1,n}, \dots, \hat{\theta}'_{\mathbb{M}_d,n})'$
- $\theta_n^* = (\theta^*_{\mathbb{M}_1,n}, \dots, \theta^*_{\mathbb{M}_d,n})'$
- Let $k = \sum_{j=1}^d m(\mathbb{M}_{j,n})$, be the dimension of $\hat{\theta}_n$

- Let $r_n = \hat{\theta}_n - \theta_n^*$
- Let $S_n = \mathbb{V}C_n(r_n)$
- Let d_w be a distance generating the topology of weak convergence for distributions on an Euclidean space
- Let $\text{corr}(\Sigma) = \text{diag}(\Sigma)^{-1/2} \Sigma \text{diag}(\Sigma)^{-1/2}$, where $\text{diag}(\Sigma)$ is obtained by setting the off-diagonal elements of Σ to 0.

Lemma

Under some conditions, we have, with $\mathbb{P}_n \circ f$ the push-forward measure of a function f under \mathbb{P}_n ,

$$d_w \left(\mathbb{P}_n \circ \left[\text{diag}(S_n)^{-1/2} \left(\hat{\theta}_n - \theta_n^* \right) \right], N(0, \text{corr}(S_n)) \right) \rightarrow 0$$

- For $\alpha \in (0, 1)$ and for a covariance matrix Γ , let $K_{1-\alpha}(\Gamma)$ be the $1 - \alpha$ -quantile of $\|Z\|_\infty$ for $Z \sim N(0, \Gamma)$
- For $\mathbb{M} = \mathbb{M}_{i,n} \in \mathbb{M}_n$ and $j \in \{1, \dots, m(\mathbb{M})\}$ let

$$j \star \mathbb{M} := \sum_{l=1}^{i-1} m(\mathbb{M}_{l,n}) + j,$$

($j \star \mathbb{M}$ is the index of $(\theta_{\mathbb{M},n}^*)'_j$ in $(\theta_{\mathbb{M}_1,n}^*, \dots, \theta_{\mathbb{M}_d,n}^*)'$)

Confidence intervals based on a consistent estimator of the asymptotic covariance matrix

Let $\alpha \in (0, 1)$. Let \hat{S}_n be so that, with $\|\cdot\|$ the largest singular value of A ,

$$\|\text{corr}(\hat{S}_n) - \text{corr}(\text{VC}_n(r_n))\| + \|\text{diag}(\text{VC}_n(r_n))^{-1} \text{diag}(\hat{S}_n) - I_k\| \rightarrow_p 0$$

Consider, for $\mathbb{M} \in \mathbb{M}_n$ and $j = 1, \dots, m(\mathbb{M})$ the confidence interval

$$\text{CI}_{1-\alpha, \mathbb{M}}^{(j), \text{est}} = \hat{\theta}_{\mathbb{M}, n}^{(j)} \pm \sqrt{[\hat{S}_n]_{j^* \mathbb{M}}} K_{1-\alpha} (\text{corr}(\hat{S}_n))$$

Theorem

Then, $\mathbb{P}_n \left(\left[\theta_{\hat{\mathbb{M}}_n, n}^* \right]_j \in \text{CI}_{1-\alpha, \mathbb{M}}^{(j), \text{est}} \text{ for all } \mathbb{M} \in \mathbb{M}_n \text{ and } j = 1, \dots, m(\mathbb{M}) \right)$ goes to $1 - \alpha$ as $n \rightarrow \infty$. In particular, for any model selection procedure $\hat{\mathbb{M}}_n$, we have

$$\liminf_{n \rightarrow \infty} \mathbb{P}_n \left(\left[\theta_{\hat{\mathbb{M}}_n, n}^* \right]_j \in \text{CI}_{1-\alpha, \hat{\mathbb{M}}_n}^{(j), \text{est}} \text{ for all } j = 1, \dots, m(\hat{\mathbb{M}}_n) \right) \geq 1 - \alpha$$

Confidence intervals based on a conservative estimator of the asymptotic covariance matrix

- When the models are misspecified it may not be possible to estimate $\mathbb{V}C_n(r_n)$ consistently
- We show how to **overestimate** the diagonal components of $\mathbb{V}C_n(r_n)$
- This is based on overestimating $\mathbb{V}(y_{i,n})$ based on

$$\mathbb{V}(y_{i,n}) \leq \mathbb{E}((y_{i,n} - \hat{y}_{i,n})^2)$$

where $\hat{y}_{i,n}$ is obtained from a misspecified model \mathbb{M}

- Also there exist upper-bounds of $K_{1-\alpha}(\text{corr}(S_n))$ (see [Berk et al 2013](#), [Bachoc Leeb Pötscher 2019](#))

⇒ We obtain the same asymptotic guarantee as before with more conservative confidence intervals

- We have seen a general method that can be applied to specific situations on a case by case basis
- Need uniform central limit theorems for fixed models in misspecified cases (sandwich rule)
- Need to consistently overestimate variances
- In the paper, we provide applications to
 - homoscedastic linear models to homoscedastic data
 - heteroscedastic linear models to heteroscedastic data
 - Binary regression models to binary data

- 1 The post-model selection inference setting
- 2 The confidence intervals for Gaussian homoscedastic linear models
- 3 The confidence intervals for more general situations
- 4 Some simulation results**

Some simulation results

In a Monte Carlo simulation (1000 repetitions) for logistic regression ($p = 10, n = 30, 100$), we compare

- CI coverage for a nominal level at 0.9 (cov. 0.9)
- CI median length (med.)
- CI 90% quantile length (qua.)

for

- our post-selection inference CI (P)
- the CI by [Taylor and Tibshirani, 2017](#), specific to the lasso (L)
- the naive CI that ignores the presence of model selection (N)

model selector	cov. 0.9			med.			qua.		
	P	L	N	P	L	N	P	L	N
lasso (1)	0.99	0.89	0.84	4.26	7.44	2.09	6.97	43.33	3.42
lasso (2)	1.00	0.85	0.68	1.63	2.31	0.74	1.90	13.52	0.84
lasso (3)	1.00	0.25	0.98	2.22	1.23	1.01	2.83	3.50	1.24
sig. hun.	0.95		0.39	4.40		2.63	6.22		3.63

Some simulation results in high dimension 1

In a Monte Carlo simulation (1000 repetitions) for homoscedastic linear models ($p = 1000, n = 50$)

- The model selector is **forward stepwise**

we compare

- CI coverage for a nominal level at 0.9 (cov. 0.9)
- CI median length (med.)
- CI 90% quantile length (qua.)

for

- our post-selection inference CI (P)
- the CI by [Tibshirani et al. 2017](#), specific to forward-stepwise (FS)
- the naive CI that ignores the presence of model selection (N)


Some simulation results in high dimension 2

	cov.	Step 1 med.	qua.	cov.	Step 2 med.	qua.	cov.	Step 3 med.	qua.	Simult. cov.
P	0.99	8.33	9.38	1.00	10.39	12.73	1.00	11.49	14.35	0.99
FS	0.94	11.66	55.76	0.88	786.92	Inf	0.90	1754.00	Inf	0.77
N	0.58	3.54	3.98	0.49	3.33	4.08	0.45	3.22	4.03	0.08
P	0.91	7.24	8.07	1.00	9.34	12.15	1.00	10.36	13.68	0.91
FS	0.93	15.15	72.67	0.88	752.74	Inf	0.90	1582.32	Inf	0.76
N	0.00	3.07	3.43	0.12	3.00	3.90	0.19	2.91	3.84	0.00

Remark

- Top 3 rows : design matrix X has independent columns
- Bottom 3 rows : design matrix X has correlated columns
- The CI's P and FS use the knowledge that k variables are selected at step k

The paper :

 **F. Bachoc, D. Preinerstorfer, L. Steinberger. Uniformly valid confidence intervals post-model-selection, <https://arxiv.org/abs/1611.01043>
Annals of statistics, forthcoming**

- We provide general asymptotic post-model selection confidence intervals
 - ▷ general results
 - ▷ applications to homoscedastic and heteroscedastic linear models and to binary regression
- for misspecified models
- with encouraging numerical behavior
- **Open questions** : high-dimensional asymptotics, computational aspects

Thank you for your attention !