Uniformly valid confidence intervals post-model-selection

François Bachoc, David Preinerstorfer and Lukas Steinberger

Institut de Mathématiques de Toulouse
Aarhus University
University of Freiburg

Journées de Statistique
Avignon, May 2017

Talk outline

Data and models

Data :

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

Models :

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

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

Parameters and estimators

Parameters :

- We define for each model $\mathbb{M} \in \mathsf{M}_n$ 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, $\mathbb{M} \in \mathsf{M}_n$ is a set of distributions parameterized by $\theta \in \mathbb{R}^{m(\mathbb{M})}$, 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 $\mathbb{M} \in \mathsf{M}_n$, an estimator $\hat{\theta}_{\mathbb{M},n}$ of the optimal parameter $\theta^*_{\mathbb{M},n}$
- The estimator $\hat{\theta}_{\mathbb{M},n}$ is a measurable function from $\mathbb{R}^{n\times\ell}$ to $\mathbb{R}^{m(\mathbb{M})}$

Post-model selection inference

Model selection :

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

In the literature :

- Van der Geer et al. 2014, AoS address linear regression, with the lasso model selector. They assume a well-specified (sparse) model and construct confidence intervals for the regression coefficients
- Lee et al. 2016, AoS address Gaussian linear regression with the lasso model selector. They consider misspecified models and construct confidence intervals for the optimal coefficients
- Berk et al 2013, AoS address Gaussian linear regression with any model selector. They also consider misspecified models.
- Taylor and Tibschirani 2017, CJoS address misspecified generalized linear models with l^1 penalized likelihood

The case of homoscedastic linear model

Setting :

- $l = 1$ and $y_{1,1}, ..., y_{n,n}$ have identical variance
- A model $\mathbb{M}$ is given by a subset of $\{1, ..., p\}$ and corresponds to extracting columns of a $n \times p$ design matrix X_n
- $\hat{\theta}_{\mathbb{M},n}$ is the restricted least square estimator in a linear model

Confidence intervals :

- Berk et al 2013, AoS observe that the vector $\{\hat{\theta}_{\mathbb{M},n} - \theta^*_{\mathbb{M},n}\}_{\mathbb{M} \in \mathrm{M}_n}$ is Gaussian
- They use a worst case approach (in terms of the selected model) and obtain a confidence interval $\mathrm{CI}^{(j)}_{1-\alpha,\hat{\mathbb{M}}_n}$, for component j of $\theta^*_{\hat{\mathbb{M}}_n,n}$ satisfying

$$\mathbb{P}_n \left(\theta^{*(j)}_{\hat{\mathbb{M}}_n,n} \in \mathrm{CI}^{(j)}_{1-\alpha,\hat{\mathbb{M}}_n} \text{ for all } j = 1, \ldots, m(\hat{\mathbb{M}}_n) \right) \geq 1 - \alpha$$

Universality :

- The coverage guarantee holds for any model selector $\hat{\mathbb{M}}_n$. Berk et al. hence speak of universally valid confidence intervals
- This universality is particularly beneficial when the statistician has limited control on the model selection procedure : informal , cost-driven...

Main idea and notation

Main idea :

- We aim at showing a joint asymptotic normality of $\{\hat{\theta}_{\mathbb{M},n} - \theta^*_{\mathbb{M},n}\}_{\mathbb{M} \in \mathsf{M}_n}$
- 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 :

- $\hat{\theta}_n = (\hat{\theta}'_{\mathbb{M}_1,n}, \ldots, \hat{\theta}'_{\mathbb{M}_d,n})'$
- $\theta^*_n = (\theta^{*'}_{\mathbb{M}_1,n}, \ldots, \theta^{*'}_{\mathbb{M}_d,n})'$
- Let $k = \sum_{j=1}^d m(\mathbb{M}_{j,n})$, be the dimension of $\hat{\theta}_n$
- Let $\mathbb{E}_n$, $\mathbb{V}_n$, and $\mathbb{VC}_n$, be the mean, the variance and the covariance matrix under $\mathbb{P}_n$
- Similarly, define$\mathbb{E}_{i,n}$, $\mathbb{V}_{i,n}$, and $\mathbb{VC}_{i,n}$ for $\mathbb{P}_{i,n}$ and $\mathbb{E}$, $\mathbb{V}$, and $\mathbb{VC}$ for $\mathbb{P}$

General assumption

There exist Borel-measurable functions $g_{i,n} : \mathbb{R}^{1\times \ell} \to \mathbb{R}^k$ for $i = 1, \ldots, n$, possibly depending on θ_n^*, so that

$$\hat{\theta}_n(y_n) - \theta_n^* \quad = \quad \sum_{i=1}^{n} g_{i,n}(y_{i,n}) + \Delta_n(y_n),$$

where, with $r_n(y_n) := \sum_{i=1}^n g_{i,n}(y_{i,n})$, we have for all $i \in \{1, \ldots, n\}$ and for all $j \in \{1, \ldots, k\}$ that

$$\mathbb{E}_{i,n}\left(g_{i,n}^{(j)}\right) = 0 \quad \text{and} \quad 0 < \mathbb{V}_n\left(r_n^{(j)}\right) < \infty,$$

and for all $j \in \{1, \ldots, k\}$ we have, with $\{.\}$ the indicator function,

$$\mathbb{V}_n^{-1}\left(r_n^{(j)}\right) \sum_{i=1}^{n} \int_{\mathbb{R}^{1\times\ell}} \left[g_{i,n}^{(j)}\right]^2 \left\{|g_{i,n}^{(j)}| \geq \varepsilon \mathbb{V}_n^{\frac{1}{2}}(r_n^{(j)})\right\} d\mathbb{P}_{i,n} \to 0 \text{ for all } \varepsilon > 0,$$

and

$$\mathbb{P}_n\left(\left|\mathbb{V}_n^{-1/2}\left(r_n^{(j)}\right)\Delta_n^{(j)}\right| \geq \varepsilon\right) \to 0 \text{ for all } \varepsilon > 0$$

Joint Asymptotic normality

- Let
$$S_n(y_n) := \sum_{i=1}^{n} g_{i,n}(y_{i,n}) g'_{i,n}(y_{i,n})$$
- Let $A^\dagger$ be the Moore-Penrose inverse of a square matrix A
- Let $A^{\dagger/2} = [A^\dagger]^{1/2}$
- Let d_w be a distance generating the topology of weak convergence for distributions on an Euclidean space
- Let $\mathrm{corr}(\Sigma) = \mathrm{diag}(\Sigma)^{\dagger/2} \Sigma \, \mathrm{diag}(\Sigma)^{\dagger/2}$, where $\mathrm{diag}(\Sigma)$ is obtained by setting the off-diagonal elements of Σ to 0.

Lemma

Under the previous condition, for $\varepsilon > 0$ we have, with $\mathbb{P}_n \circ f$ the push-forward measure of a function f under $\mathbb{P}_n$,

$$\mathbb{P}_n \left(d_w \left(\mathbb{P}_n \circ \left[\mathrm{diag}(S_n)^{\dagger/2} \left(\hat{\theta}_n - \theta_n^* \right) \right], N(0, \mathrm{corr}(S_n)) \right) \geq \varepsilon \right) \to 0,$$

and this continues to hold when replacing S_n by $\mathbb{VC}_n(r_n)$

Some notation

- 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 $\mathbb{M} = \mathbb{M}_{i,n} \in \mathsf{M}_n$ and $j \in \{1,\dots,m(\mathbb{M})\}$ let

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

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

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

Theorem

Let $\alpha \in (0,1)$. Let $\hat{S}_n : \mathbb{R}^{n \times \ell} \to \mathbb{R}^{k \times k}$ be so that for all $\varepsilon > 0$, with $||.||$ the largest singular value of A,

$$\mathbb{P}_n \left(\| \operatorname{corr}(\hat{S}_n) - \operatorname{corr}\left(\mathbb{VC}_n(r_n) \right) \| + \| \operatorname{diag}(\mathbb{VC}_n(r_n))^{-1} \operatorname{diag}(\hat{S}_n) - I_k \| \geq \varepsilon \right)$$

goes to 0*. Consider, for $\mathbb{M} \in \mathrm{M}_n$ and $j = 1, \ldots, m(\mathbb{M})$ the confidence interval*

$$\mathrm{CI}_{1-\alpha,\mathbb{M}}^{(j),\mathrm{est}} \quad = \quad \hat{\theta}_{\mathbb{M},n}^{(j)} \pm \sqrt{[\hat{S}_n]_{j\star\mathbb{M}}}\, K_{1-\alpha} \left(\operatorname{corr}(\hat{S}_n) \right)$$

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

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

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

- When the models are misspecified it may not be possible to estimate $\mathbb{VC}_n(r_n)$ consistently
- We show how to overestimate the diagonal components of $\mathbb{VC}_n(r_n)$
- 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 $K_{1-\alpha}\left(\mathrm{corr}(\hat{S}_n)\right)$ (see Berk et al 2013, Bachoc Leeb Pötscher 2017+)

Notation

Uniform asymptotic :

- We now provide asymptotic results that are uniform over sets $\mathbf{P}_n$ of possible distributions $\mathbb{P}_n$ for the observations

$\Rightarrow$ This is why we worked with generic triangular arrays before

Notation :

- For a Borel set T of $\mathbb{R}$, we let $M(T^n)$ be the set of probability measures on $T^n = \times_{i=1}^n T \subseteq \mathbb{R}^n$
- The mean vector of Q in $M(T^n)$ is written $\mu(Q)$
- For $Q \in M(T^1)$ and for $0 < q < \infty$, we write $m_q(Q)$ for the q-th absolute centered moment of Q
- We let $\bigotimes_{i=1}^n M(T)$ be the set of product measures on $M(T^n)$.
- For $Q \in \bigotimes_{i=1}^n M(T)$ we let $Q = \bigotimes_{i=1}^n Q_i$.

Homoscedastic linear models

Set of possible distributions :

$$\mathbf{P}_n^{(\mathrm{lm})}(\delta, \tau) := \left\{ Q \in \bigotimes_{i=1}^{n} M(\mathbb{R}) : \begin{array}{l} 0 < m_2(Q_1) = \ldots = m_2(Q_n) < \infty \\ \frac{\max_{i=1,\ldots,n} m_{2+\delta}(Q_i)^{\frac{2}{2+\delta}}}{m_2(Q_1)} \leq \tau \end{array} \right\},$$

for fixed $\delta > 0$ and $\tau > 1$

Models :

- We consider a fixed observed $n \times p$ design matrix X_n (p fixed)
- It is known that the observations have the same variance
- Each model $\mathbb{M}_{i,n}$ is defined by a subset M_i of $\{1, \ldots, p\}$
- We let $X_n[M]$ be the matrix obtained by deleting the columns of X_n which indices are not in $M \subset \{1, ..., p\}$
- We let, for $j \in \{1, \ldots, d\}$,

$$\mathbb{M}_{j,n} = \left\{ Q \in \bigotimes_{i=1}^{n} M(\mathbb{R}) : \begin{array}{l} 0 < m_2(Q_1) = \ldots = m_2(Q_n) < \infty \\ \mu(Q) \in \mathrm{span}(X_n[M_j]) \end{array} \right\}$$

Confidence intervals

For a model $\mathbb{M} \in \mathrm{M}_n$ with set of variables $M \subset \{1, ..., p\}$, the optimal parameter is

$$\beta^*_{\mathbb{M},n} = \beta^*_{\mathbb{M},n}(\mathbb{P}_n) = \left(X_n[M]' X_n[M]\right)^{-1} X_n[M]' \mu(\mathbb{P}_n)$$

It satisfies, for all $\beta \in \mathbb{R}^{m(\mathbb{M})}$,

$$||\mu(\mathbb{P}_n) - X_n[M]\beta^*_{\mathbb{M},n}|| \leq ||\mu(\mathbb{P}_n) - X_n[M]\beta||$$

$\Longrightarrow$ We obtain the same asymptotic coverage guarantees as in the general case

Heteroscedastic linear models

The principle is the same. The set of possible observation distributions is

$$\mathbf{P}_n^{(\mathrm{het})}(\delta,\tau) := \left\{ Q \in \bigotimes_{i=1}^{n} M(\mathbb{R}) : \begin{array}{l} 0 < m_2(Q_i) < \infty \text{ for } i = 1,\ldots,n \\ \frac{\max_{i=1,\ldots,n} m_{2+\delta}(Q_i)^{\frac{2}{2+\delta}}}{\min_{i=1,\ldots,n} m_2(Q_i)} \leq \tau \end{array} \right\}$$

and the set of models $\mathsf{M}_n = \{\mathbb{M}_{j,n} : j = 1,\ldots,d\}$ is given by

$$\mathbb{M}_{j,n} = \left\{ Q \in \bigotimes_{i=1}^{n} M(\mathbb{R}) : \begin{array}{l} 0 < m_2(Q_i) < \infty \text{ for } i = 1,\ldots,n \\ \mu(Q) \in \mathrm{span}(X_n[M_j]) \end{array} \right\}$$

The optimal parameters are the same as before

Binary regression

- In binary regression, the set of observation distributions depends on $\tau > 0$ and is given by

$$\mathbf{P}_n^{(\mathrm{bin})}(\tau) := \left\{ Q \in \bigotimes_{i=1}^{n} M(\{0,1\}) : Q_i(\{0\}) Q_i(\{1\}) \geq \tau \; \forall i = 1, \ldots, n \right\}$$

- We consider the set of models $\mathsf{M}_n = \left\{ \mathbb{M}_{(j_1,j_2),n} : j_1 \in \{1, \ldots, d_1\}, j_2 \in \{1, \ldots, d_2\} \right\}$, given by
  - response functions $h_1, ..., h_{d_1} : \mathbb{R} \to [0,1]$
  - subsets $M_1, ..., M_{d_2}$ of $\{1, ..., p\}$
- Let $X_{i,n}$ be line i of X_n
- Then the submodels are defined by

$$\mathbb{M}_{(j_1,j_2),n} = \left\{ Q \in \bigotimes_{i=1}^{n} M(\{0,1\}) : \begin{array}{l} \exists \beta \in \mathbb{R}^{|M_{j_2}|} : \forall i = 1, \ldots, n : \\ Q_i(\{1\}) = h_{j_1}(X_{i,n}[M_{j_2}]\beta) \end{array} \right\}$$

- We obtain the same asymptotic coverage guarantees as in the general case
- In the case of canonical response function (logistic regression), the confidence intervals become shorter

Conclusion

- We provide general asymptotic post-model selection confidence intervals
  - in non-Gaussian cases
  - for misspecified models
- Prospects :
  - Numerical comparison with other post-model-selection confidence intervals (ongoing)
  - What can be done in the high-dimensional asymptotic case ?

The paper :

F. Bachoc, D. Preinerstorfer, L. Steinberger. Uniformly valid confidence intervals post-model-selection, **`https://arxiv.org/abs/1611.01043`**

Thank you for your attention !