KU LEUVEN

Asymptotic post-selection inference after model selection by the Akaike information criterion.

Gerda Claeskens

ORSTAT and Leuven Statistics Research Center, KU Leuven, Belgium

WHOA-PSI-2017, St. Louis, 12–14 August 2017

Confidence intervals post-AIC selection

Joint work with **Ali Charkhi** (KU Leuven)

Model selection by AIC

- Model selection by Akaike's information criterion

$$\text{AIC}(M_k) = -2 \log \text{Likelihood}_{M_k}(\widehat{\beta}_{M_k}) + 2p_{M_k}$$

  Select model $\hat{M}_{\text{aic}}$ for which $\text{AIC}(\widehat{M}_{\text{aic}})$ is the smallest amongst $\text{AIC}(M_k), k = 1, \ldots, K$

- Estimation in the selected model:

$$\mu(\widehat{\beta}_{\hat{M}_{\text{aic}}}) = \sum_{k=1}^{K} \mu(\widehat{\beta}_{M_k}) \cdot I(M_k = \widehat{M}_{\text{aic}})$$

Selection is a special case of model averaging:

- Weights are only 0 or 1
- Weights are random, since data-driven.
- Randomness of weights influences further inference

Confidence regions after AIC selection

In selected model $\widehat{M}_{\rm aic}$, estimate focus μ by

$$\widehat{\mu}(\widehat{M}_{\rm aic}) = \sum_{k=0}^{K} \widehat{\mu}(M_k) \cdot I(M_k = \widehat{M}_{\rm aic}).$$

Confidence regions for μ after selection?
Naive interval: pretend that the model has been given beforehand. Ignore randomness $\widehat{M}_{\rm aic} = M_{\rm aic}$.

$$\hat{\mu}(M_{\rm aic}) \pm 1.96 \cdot \text{standard error}(\hat{\mu}(M_{\rm aic}))$$

Frequentist model averaging necessary to overcome low coverage of using naive method after AIC-selection.

(Hjort and Claeskens, 2003)

Some recent literature on correct inference

Berk, Brown, Buja, Zhang, Zhao (2013): valid confidence intervals irrespective of the selection procedure. No overall true values for the parameters.

Bachoc, Leeb, Pötscher (2015): generalized above method to prediction intervals.

Efron (2014): calculate the variance of the bagging estimator $\Rightarrow$ not based on the selected model for the original data.

Lee, Sun, Sun, Taylor (2016): obtained exact post-selection inference for lasso model selection for a given value of λ.

Schneider (2016): coverage of intervals based on thresholding estimators in high-dimensional linear regression models.

Post-AIC-selection in nested models

Classical (easy) setting to start off with:

- $\mathcal{M}_{\text{nest}} = \{M_k; M_1 \subset M_2 \subset \ldots \subset M_K\}$
- Maximum likelihood estimation in each model $\widehat{\boldsymbol{\theta}}_{(k)} = (\widehat{\boldsymbol{\theta}'}^t_{(k)}, \mathbf{0}^t_{K-k})^t$.
- Minimal true model $M_{p_0} \in \mathcal{M}_{\text{nest}}$:
  Models with indices $k < p_0$ are under-parametrized,
  Models with $k > p_0$ are over-parametrized.
- $\text{AIC}(M_k) = -2\ell_n(\widehat{\boldsymbol{\theta}}_{(k)}) + 2(a + k)$

Select $\widehat{p}_{\text{aic}}$ if and only if $\text{AIC}(M_{\widehat{p}_{\text{aic}}}) \leq \text{AIC}(M_k)$ for all $k = 1, \ldots, K$

Asymptotics of AIC selection

AIC overselects (Nested models: Woodroofe 1982, arc-sine laws)

for $n \to \infty$, $\widehat{p}_{\text{aic}} \in \{p_0, p_0+1, \ldots, K\}$,

Minimize AIC $\Leftrightarrow$ maximize AIC*

$$\text{AIC}^*(M_k) = 2\{\ell_n(\widehat{\boldsymbol{\theta}}_{(k)}) - \ell_n(\boldsymbol{\vartheta})\} - 2k = 2\ell^*_{n,k} - 2k.$$

Joint convergence

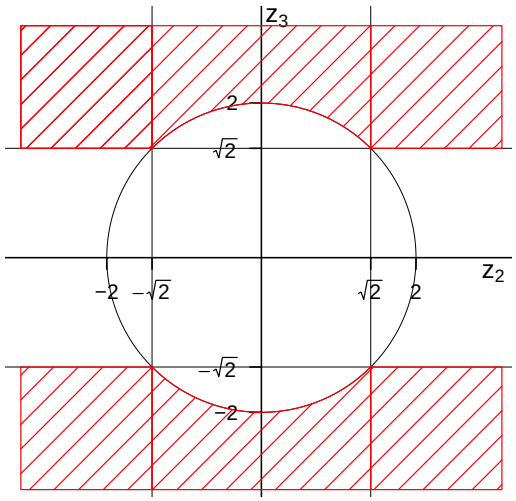
$$2(\ell^*_{n,p_0}, \ldots, \ell^*_{n,K}) \xrightarrow{d} (\underbrace{\sum_{i=1}^{a+p_0} Z_i^2}_{\chi^2_{a+p_0}}, \ldots, \underbrace{\sum_{i=1}^{a+K} Z_i^2}_{\chi^2_{a+K}})$$

with $Z_1, \ldots, Z_{a+K} \overset{i.i.d.}{\sim} N(0,1)$

AIC selection region – Example

Three models: $M_1 : \{\theta_1\}$; $M_2 : \{\theta_1, \theta_2\}$; $M_3 : \{\theta_1, \theta_2, \theta_3\}$

$a = 1$, $K = 2$, take M_1 smallest true model, M_3 selected, $\widehat{p}_{\text{aic}} = 2$



$$\begin{cases} \text{AIC}^*(M_3) > \text{AIC}^*(M_2) \\ \text{AIC}^*(M_3) > \text{AIC}^*(M_1) \end{cases}$$

Asymptotic representation:

$$\begin{cases} Z_1^2 + Z_2^2 + Z_3^2 - 6 > Z_1^2 + Z_2^2 - 4 \\ Z_1^2 + Z_2^2 + Z_3^2 - 6 > Z_1^2 - 2 \end{cases}$$

$$\mathcal{A}_2(\mathcal{M}_{\text{nest}}) = \{\boldsymbol{z} \in \mathbb{R}^3 : z_3^2 > 2, z_2^2 + z_3^2 > 4\}.$$

Asymptotic distribution

$\boldsymbol{J}_p$: Fisher information for model M_p.
$\widetilde{\boldsymbol{\nu}}_{(p)} = (\nu_1, \ldots, \nu_{a+p})^t$: sub-vector of $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_{a+K})^t$ for model M_p.
$\boldsymbol{\vartheta}$ true value

Post-AIC distribution for nested models

For $\mathcal{M}_{\text{nest}}$ with p_0 denoting the true model order

$$\begin{aligned} F_p(\boldsymbol{t}) &= \lim_{n\to\infty} P(n^{1/2}(\widehat{\boldsymbol{\theta}}_{(p)} - \boldsymbol{\vartheta}) \leq \boldsymbol{t} \mid \widehat{p}_{\text{aic}} = p, \mathcal{M}_{\text{nest}}) \\ &= P\{\boldsymbol{J}_p^{-1/2}\widetilde{\boldsymbol{Z}}_{(p)} \leq \widetilde{\boldsymbol{t}}_{(p)} \mid \widetilde{\boldsymbol{Z}}_{(p)} \in \mathcal{A}_p^{(s)}(\mathcal{M}_{\text{nest}})\} \cdot I(\boldsymbol{t} \in \mathcal{T}_p), \end{aligned}$$

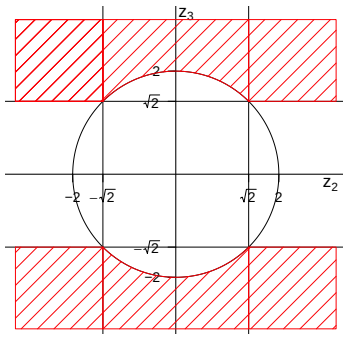
where $\mathcal{A}_p^{(s)}(\mathcal{M}_{\text{nest}}) = \{\widetilde{\boldsymbol{z}}_{(p)} \in \mathbb{R}^{a+p} : \bigcap_{j=p_0+1,\ldots,p}[\sum_{i=j}^{p}(z_{a+i}^2 - 2) > 0]\}$ and $\mathcal{T}_p = \mathbb{R}^{a+p} \times (\mathbb{R}^+)^{K-p}$.

Simplified region for nested models due to independence of $\widetilde{\boldsymbol{Z}}_{(p)}$ and $(Z_{p+1}, \ldots, Z_K)$.

Asymptotic density

Truncated multivariate normal density

Denote $\phi_p(\cdot|\mathcal{A};\Sigma)$ the density of $\Sigma^{-1/2}\widetilde{\boldsymbol{Z}}_{(p)}$ where $\widetilde{\boldsymbol{Z}}_{(p)} \sim N_{a+p}(\boldsymbol{0}, \boldsymbol{I}_{a+p})$ is truncated such that $\widetilde{\boldsymbol{Z}}_{(p)} \in \mathcal{A}$.



Corollary: Post-AIC density for nested models

The limiting density function of $n^{1/2}(\widehat{\boldsymbol{\theta}}_{(\widehat{p}_{\mathrm{aic}})} - \boldsymbol{\vartheta})$ conditional on the AIC-selection with $\widehat{p}_{\mathrm{aic}} = p$ from the set of nested models $\mathcal{M}_{\mathrm{nest}}$ is

$$f_p(\boldsymbol{t}) = \phi_p(\widetilde{\boldsymbol{t}}_{(p)}|\mathcal{A}_p^{(s)}(\mathcal{M}_{\mathrm{nest}}); \boldsymbol{J}_p^{-1}) \cdot I(\boldsymbol{t} \in \mathcal{T}_p).$$

When the true model is selected, $\widehat{p}_{\mathrm{aic}} = p_0$,
$f_{p_0}(\boldsymbol{t}) = \phi_{p_0}(\widetilde{\boldsymbol{t}}_{(p_0)}) \cdot I(\boldsymbol{t} \in \mathcal{T}_p)$

Exact calculations for nested models

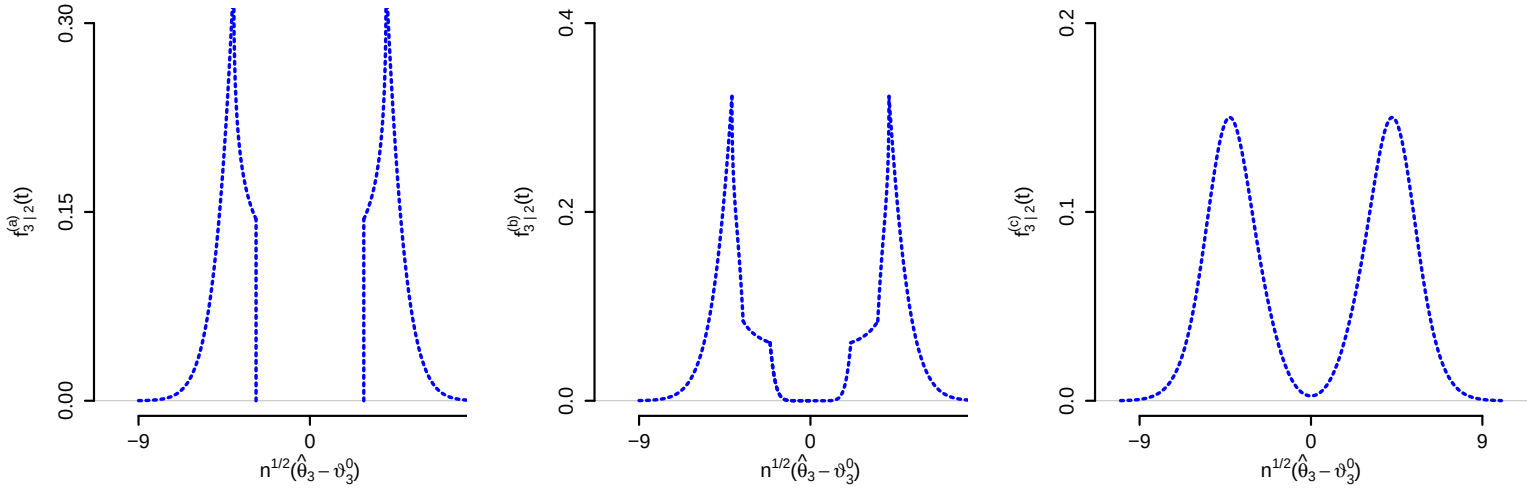
True value for parameters: $\vartheta = (\vartheta_1, 0, 0)^t \hookrightarrow M_1$ true.
Models set: $\mathcal{M}_{\text{nest}} = \{M_1, M_2, M_3\}$.
Selected model: $M_3 \Longrightarrow \mathcal{A}_3 = \{(z_1, z_2, z_3) : z_3^2 > 2, z_3^2 + z_2^2 > 4\}$.
Selection probability: $P(\boldsymbol{Z} \in \mathcal{A}_3) = 0.08$.

$\boldsymbol{J}^{-1/2}(\boldsymbol{\vartheta})$:

$$(a) \begin{pmatrix} 1.0 & 0 & 0 \\ 0 & 2.0 & 0 \\ 0 & 0 & 2.0 \end{pmatrix}, \quad (b) \begin{pmatrix} 1.0 & 0 & 0 \\ 0 & 2.0 & 0.5 \\ 0 & 0.5 & 2.0 \end{pmatrix}, \quad (c) \begin{pmatrix} 1.0 & 0.9 & 0.9 \\ 0.9 & 2.0 & 0.5 \\ 0.9 & 0.5 & 2.0 \end{pmatrix}.$$

Exact calculations for nested models

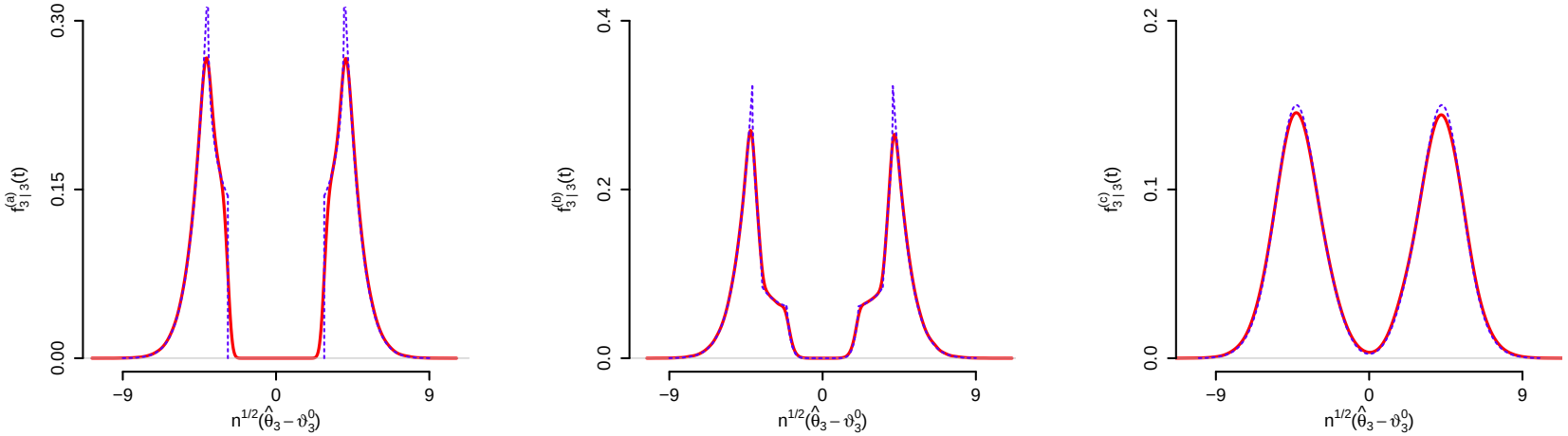
Exact marginal asymptotic distributions of $\sqrt{n}(\widehat{\theta}_3 - \vartheta_3)$ when M_1 is true model and M_3 is selected, for different Fisher information matrices.



Simulation based inference

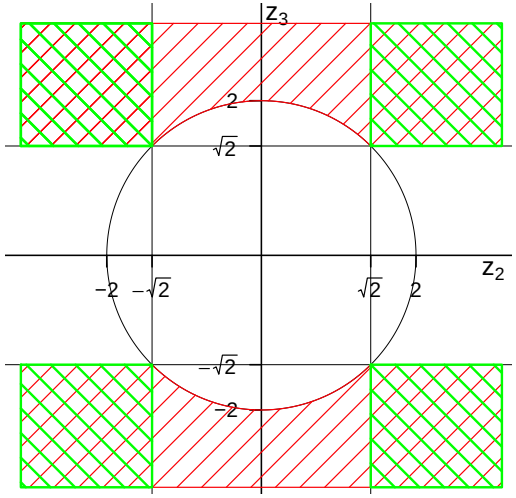
Sample from a multivariate normal distribution subject to quadratic constraints. (Pakman and Paninski 2014)

(a) (b) (c)



simulated quantile	coverage	χ^2 quantile	coverage
12.17	0.949	7.81	0.826

Inference depends on the set of models



Nested models:

$M_1 : \{\theta_1\}$; $M_2 : \{\theta_1, \theta_2\}$; $M_3 : \{\theta_1, \theta_2, \theta_3\}$

AIC selects M_3:

AIC$(M_3) >$ AIC(M_1) and AIC$(M_3) >$ AIC(M_2)

$\mathcal{A}_M(\mathcal{M}_{\text{nest}}) = \{\boldsymbol{z} \in \mathbb{R}^3 : z_3^2 > 2, z_2^2 + z_3^2 > 4\}$.

$\mathcal{M}_{\text{all}} =$ All subsets of $M_3 : \{\theta_1, \theta_2, \theta_3\}$

AIC selects M_3:

AIC$(M_3) >$ AIC(M') for all models $M' \neq M_3 \in \mathcal{M}$

$$\mathcal{A}_M(\mathcal{M}_{\text{all}}) = \{\boldsymbol{z} \in \mathbb{R}^3 : z_2^2 > 2, z_3^2 > 2, z_2^2 + z_3^2 > 4\}$$

Different area, different conditioning

$\Rightarrow$ different distribution
$\Rightarrow$ different quantiles
$\Rightarrow$ different confidence intervals

General likelihood models, any model set

For $M_{aic} = M \in \mathcal{M}_O$, overparametrized model, the confidence region

$$C(q_\alpha) = \left\{ \boldsymbol{\theta} \in \mathbb{R}^{a+K} : n(\widehat{\boldsymbol{\theta}}_{(M)} - \widetilde{\boldsymbol{\theta}}_{(M)})^t \boldsymbol{J}_M (\widehat{\boldsymbol{\theta}}_{(M)} - \widetilde{\boldsymbol{\theta}}_{(M)}) \leq q_\alpha \right\},$$

where q_α is determined by solving

$$\frac{P\left(\{\sum_{i \in M} Z_i^2 \leq q_\alpha\} \cap \boldsymbol{Z} \in \mathcal{A}_M(\mathcal{M}_{\text{arb}} \cap \mathcal{M}_O)\}\right)}{P(\boldsymbol{Z} \in \mathcal{A}_M(\mathcal{M}_{\text{arb}} \cap \mathcal{M}_O))} = 1 - \alpha.$$

Uniform result

Under some assumptions, incl. compact parameter space.

$$\lim_{n \to \infty} \inf_{\boldsymbol{\vartheta} \in \Theta} P_\vartheta \{\boldsymbol{\vartheta} \in C(q_\alpha) \mid M_{aic} \in \mathcal{M}_O\} = 1 - \alpha.$$

When $\mathcal{A}_M(\mathcal{M}_{\text{arb}})$ replaces $\mathcal{A}_M(\mathcal{M}_{\text{arb}} \cap \mathcal{M}_O)$) to obtain $\tilde{q}_\alpha$, $\lim_{n \to \infty} \inf_{\boldsymbol{\vartheta} \in \Theta} P\{\boldsymbol{\vartheta} \in C(\tilde{q}_\alpha) \mid M_{aic} \in \mathcal{M}_O\} \geq 1 - \alpha$.

About the uniformity result

- Limitation of AIC: selection of an overspecified model does not happen in a uniform way (Leeb, Pötscher, 2003).
- An overspecified model is needed. Hence result cannot be strengthened for AIC under this setup.
- For underparametrized model: need to work with pseudo-true values.
- 'Targets' differ from model to model.

All likelihood models misspecified

- Triangular array of observations $\{Y_{ni}; i = 1, \ldots, n, \text{ and } n \in \mathbb{N}_0\}$;
- True distribution of $(Y_{n1}, \ldots, Y_{nn})$ is G_n
- Estimators maximizing $\prod_{i=1}^n f_{ji}(y_i; \theta_j)$ converge to pseudo-true values $\vartheta_n^*(M_j)$, $j = 1, \ldots, K$;
- When there is an estimator $\hat{\Sigma}$ of Σ such that

$$\lim_{n\to\infty} \sup_{G_n \in \mathcal{G}_n} P_{G_n}(\|\hat{\Sigma}_n - \Sigma\| > \varepsilon) = 0,$$

then, with $\mathcal{Z}_{m'} \sim N_{m'}(0, I_{m'})$

$$\lim_{n\to\infty} \sup_{G_n \in \mathcal{G}_n} \sup_{t \in \mathbb{R}^{m'}} |P(\hat{\Sigma}_n^{-1/2} n^{-1/2}(\hat{\theta}_{n,\mathcal{M}} - \vartheta_{n,\mathcal{M}}^*) \leq t) - P(\mathcal{Z}_{m'} \leq t)| = 0.$$

- White (1994) contains conditions. Sandwich estimator might overestimate $\hookrightarrow$ conservative confidence intervals.

Assume (i) that there is a smallest model nested in all other models, (ii) similarity assumption of Vuong (1989), then construct selection region $\mathcal{A}_{M_{\text{aic}}}(\mathcal{M})$ similar as before

Uniform result under likelihood misspecification

$$\lim_{n\to\infty} \sup_{G_n \in \mathcal{G}_n} \sup_{t \in \mathbb{R}^{|M_{\text{aic}}|}} \Big| P(n^{1/2}\{\hat{\theta}(M_{\text{aic}}) - \vartheta^*(M_{\text{aic}})\} \leq t \mid M_{\text{aic}}) - P(\Sigma^{1/2} Z \leq t \mid \mathcal{A}_{M_{\text{aic}}}(\mathcal{M}))\Big| = 0$$

Define the set

$$C^*(q_\alpha) = \{\theta \in \mathbb{R}^{|M_{\text{aic}}|} : n\{\hat{\theta}(M_{\text{aic}}) - \theta(M_{\text{aic}})\}^t \Sigma_{M_{\text{aic}}}(\vartheta^*_{M_{\text{aic}}})^{-1}\{\hat{\theta}(M_{\text{aic}}) - \theta(M_{\text{aic}})\} \leq q_\alpha\},$$

Then $\lim_{n\to\infty} \sup_{G_n \in \mathcal{G}_n} \sup_{\alpha \in [0,1]} \Big| P_{G_n}\{\vartheta^*(M_{\text{aic}}) \in C^*(q_\alpha) | M_{\text{aic}}\} - (1-\alpha)\Big| = 0.$

Simulation study – Linear Regression

Model to generate data

$$Y_i = \sum_{j=1}^{10} \theta_j x_{ji} + \varepsilon_i, \quad i = 1, \ldots, n,$$

- $\varepsilon_i \sim N(0,1)$
- $\boldsymbol{\theta}^t = (\theta_1, \ldots, \theta_{10})^t = (2.25, -1.1, 2.43, -2.24, 2.5, \mathbf{0}_5^t)$.
- $x_{1i} = 1$ and $(x_{2i}, \ldots, x_{10,i})^t \sim N(\mathbf{0}_9, \boldsymbol{\Omega})$ where $\Omega_{ii} = 1$ and $\Omega_{ij} = 0.25,\ j \neq i$.

Selection by AIC from $\boldsymbol{\zeta}_{\text{all}}^i$: first i parameters are present in all models, all combinations of other parameters.

Focus model: $M = (\theta_1, \ldots, \theta_6, \theta_8)$: Run simulation until M is selected 3000 times.

Report average confidence intervals over 3000 runs and coverage probabilities.

CI for regression parameter

PoSI: valid inference irrespective of model selection method.

(Berk, Brown, Buja, Zhang, Zhao 2013 AoS)

n	method	θ_j	ζ^3_{all}	Cov.	ζ^6_{all}	Cov.
		θ_4	[-2.54, -1.94]	0.99	[-2.46, -2.02]	0.94
	PostAIC	θ_6	[-0.30, 0.31]	0.95	[-0.22, 0.22]	0.95
		θ_8	[-0.30, 0.31]	0.95	[-0.29, 0.30]	0.95
		θ_4	[-2.58, -1.90]	1.00	[-2.54, -1.94]	0.99
100	PoSI	θ_6	[-0.33, 0.34]	0.98	[-0.30, 0.30]	0.99
		θ_8	[-0.34, 0.34]	0.98	[-0.29, 0.31]	0.95
		θ_4	[-2.46, -2.02]	0.93	[-2.46, -2.02]	0.93
	Naive	θ_6	[-0.22, 0.22]	0.66	[-0.22, 0.22]	0.94
		θ_8	[-0.22, 0.22]	0.66	[-0.22, 0.23]	0.69

ζ^3_{all} : More conservative for θ_4.

ζ^6_{all} : Almost exact for all parameters.

CI for a linear combination of the parameters $x^t\theta$

PoSIp (Bachoc, Leeb, Pötscher, 2015)
Smoothed bootstrap CI (Efron, 2014)

			ζ^3_{all}		ζ^5_{all}	
σ	n	method	length	cov.	length	cov.
1	30	PostAIC	3.11	0.97	2.61	0.95
		Bootstrap	3.67	0.92	3.32	0.92
		PoSIp	4.38	1.00	4.39	1.00
	100	PostAIC	1.42	0.98	1.17	0.95
		Bootstrap	1.25	0.94	1.25	0.94
		PoSIp	1.83	1.00	1.83	1.00
3	30	PostAIC	11.76	0.98	7.82	0.94
		Bootstrap	11.46	0.92	9.95	0.92
		PoSIp	12.65	0.99	13.16	1.00
	100	PostAIC	4.25	0.98	3.50	0.95
		Bootstrap	3.77	0.94	3.74	0.94
		PoSIp	5.47	1.00	5.48	1.00

Extensions to other criteria

AIC-like criteria, overselection is exploited.

- Takeuchi's information criterion (1976)

$$\text{TIC}(M) = -2\ell_n(\hat{\boldsymbol{\theta}}(M)) + 2\text{tr}\{\boldsymbol{Q}_M(\boldsymbol{\vartheta}^*)^{-1}\boldsymbol{J}_M(\boldsymbol{\vartheta}^*)\}.$$

  Replace $|M|$ with $\text{tr}\{\boldsymbol{Q}_M(\boldsymbol{\vartheta}^*)^{-1}\boldsymbol{J}_M(\boldsymbol{\vartheta}^*)\}$

- Generalized information criterion (Konishi, Kitagawa 1996)

$$\text{GIC}(M) = -2\ell_n(\hat{\boldsymbol{\theta}}(M)) + \frac{2}{n}\sum_{i=1}^{n}\text{tr}\Big\{\text{Infl}(Y_i)\frac{\partial}{\partial\boldsymbol{\theta}_M^t}\log f(Y_i;\hat{\boldsymbol{\theta}}_M)\Big\}.$$

- Mallows's C_p for linear regression (1973)

$$C_p(M) = \hat{\sigma}^{-2}\hat{\sigma}^2(M) + 2|M| - n$$

  When $n \to \infty$, $C_p(M) - C_p(M^*) \sim \chi_q^2/q + 2q$ where $q = |M^*| - |M|$

Take home messages

- Asymptotic AIC selection regions are formed by quadratic functions.
- Need overparametrization when a true model exists.
- Stronger uniform results under misspecification (pseudo-true parameter values).
- Think carefully about the sets of models to include in a search.

Thank you!