Confidence Envelopes for Model Selection Criteria and Post-Model Selection Inference

Nadeesha Imiya Mohottige & Alex Trindade

Dept. of Mathematics & Statistics, Texas Tech University

June 2022

- **GIC:** A *generalized information criterion* (statistic) consists of the normalized log-likelihood plus a penalty term; includes AIC, BIC, etc.
- **Goal:** Estimate a high quantile of distribution of minimum GIC (confidence envelope)
- 3 cases of special relevance: IID, regression, time series
- **Methods:**
  - analytical by invoking central limit theorems
  - bootstrap (3 versions)
- **Simulations:** accuracy of coverage probabilities of confidence envelopes
- **Illustrations:** practical implementation on real datasets

Consider 14 positive-valued models

Rank	AIC	Model
1	4.9354	Gamma3
2	4.9355	Gamma
3	4.9374	Burr
4	4.9418	Weibull
5	4.9418	Weibull3
6	4.9438	Gumbel
7	4.9631	LogNormal3
8	4.9658	Lognormal
9	4.9695	Logistic3
10	4.9706	Logistic
11	4.9776	Inverse Gaussian
12	5.0532	Kappa
13	5.1380	Gpareto
14	7.1821	F

Consider all subsets with 8 predictors (only top 20 out of 256 shown)

Rank	BIC	Model
1	7.570316919	{0123458}
2	7.571444201	{012348}
3	7.57664502	{01234568}
4	7.576713773	{0123478}
5	7.576790141	{01234578}
6	7.577886861	{0123468}
7	7.579900096	{012345678}
8	7.582701966	{01234678}
9	7.586709991	{01235678}
10	7.614226865	{0123678}
11	7.619886251	{01245678}
12	7.628053092	{0124678}
13	7.628516874	{0124578}
14	7.637781395	{012458}
15	7.643862937	{0123568}
16	7.644318814	{0124568}
17	7.651154208	{0123578}
18	7.662647169	{012358}
19	7.671062354	{01345678}
20	7.672409034	{0134678}

Models for Wolfer Sunspot Numbers: annual time series of 100 values (1770–1869)

Rank	AICc	Model
1	836.6704608	$ARMA(8,0)$
2	837.8329215	$ARMA(3,5)$
3	838.1403819	$ARMA(5,3)$
4	838.2688009	$ARMA(5,5)$
5	838.844269	$ARMA(4,4)$
6	838.8467446	$ARMA(8,1)$
7	838.9254096	$ARMA(9,0)$
8	839.0495172	$ARMA(9,1)$
9	839.8700548	$ARMA(3,2)$
10	840.0848802	$ARMA(9,2)$
11	840.5555259	$ARMA(11,1)$
12	840.9585325	$ARMA(4,2)$
13	841.1186244	$ARMA(8,2)$
14	841.1439068	$ARMA(4,5)$
15	841.1898953	$ARMA(10,1)$
16	841.1954145	$ARMA(11,0)$
17	841.3298421	$ARMA(9,13)$
18	841.7005855	$ARMA(12,13)$
19	841.7189501	$ARMA(12,0)$
20	842.4678861	$ARMA(3,3)$
21	842.4705473	$ARMA(10,2)$
22	842.5017981	$ARMA(9,3)$
23	842.8640372	$ARMA(11,2)$

- Observations:: $y_1, \ldots, y_n \sim Y$, cdf $F(y)$, pdf $f(y)$.
- Candidate model set: $\mathcal{G} = \left\{ g_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d \right\}$.
- Inference on $\boldsymbol{\theta}$ for $g_{\boldsymbol{\theta}}$: maximize log-likelihood

$$\ell_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log g(y_i; \boldsymbol{\theta}), \qquad \widehat{\boldsymbol{\theta}} = \text{MLE}$$

- SLLN: empirical KL-discrepancy converges to distance $f \mapsto g_{\boldsymbol{\theta}}$:

$$\Delta_n(\widehat{\boldsymbol{\theta}}) := -\frac{1}{n}\ell_n(\widehat{\boldsymbol{\theta}}) \xrightarrow{a.s.} -\mathbb{E}_F \log g(Y; \boldsymbol{\theta}) := \Delta(\boldsymbol{\theta})$$

- $\boldsymbol{\theta}_0$: parameter that gets $g_{\boldsymbol{\theta}_0}$ as close as possible to true f (Kullback-Leibler sense):

$$\widehat{\boldsymbol{\theta}} \xrightarrow{a.s.} \boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta} \in \Theta} \Delta(\boldsymbol{\theta}),$$

- $g(y; \boldsymbol{\theta}_0)$ is best approximating model (Claeskens & Hjort, 2008)

- Define Generalized Information Criterion (GIC):

$$GIC = \Delta_n(\widehat{\boldsymbol{\theta}}) + \frac{\psi(n, d)}{n}$$

- $\psi(n, d)$ is penalty term:

$$\begin{aligned} \text{AIC:} \quad & \psi(n, d) = d. \\ \text{BIC:} \quad & \psi(n, d) = (d/2)\log(n). \\ \text{AICc:} \quad & \psi(n, d) = nd/(n - d - 1). \end{aligned}$$

- Define GIC for j-th candidate model (a statistic):

$$X_j := \text{GIC}_j := \Delta_n(\widehat{\boldsymbol{\theta}}^{(j)}) + \frac{\psi(n, d_j)}{n},$$

- Under mild regularity conditions (Linhart, 1988):

Theorem

$$\sqrt{n}\begin{pmatrix} X_j - \Delta(\boldsymbol{\theta}_0^{(j)}) \\ X_k - \Delta(\boldsymbol{\theta}_0^{(k)}) \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \Lambda\right),$$

provided $\psi(n, d) = o(\sqrt{n})$ *for each model, where* (j, k) *element of* Λ *is given by:*

$$\lambda_{jk} \quad = \quad Cov_F\left(\log g^{(j)}(Y; \boldsymbol{\theta}_0^{(j)}), \log g^{(k)}(Y; \boldsymbol{\theta}_0^{(k)})\right) \qquad (1)$$

- Entire set of K GIC values: a dependent and non-identically distributed sequence of approximately normal random variables.
- Use obvious empirical estimates for Λ.
- Approximate cdf of min GIC:

$$F_{X_{(1)}}(x) = 1 - P(X_1 > x, \ldots, X_K > x).$$

- Efficient computation via R package **mvtnorm** for up to $K = 1000$.
- **Confidence envelope for min GIC of level $(1-\alpha)100\%$:** $(1-\alpha)$ quantile of $X_{(1)}$

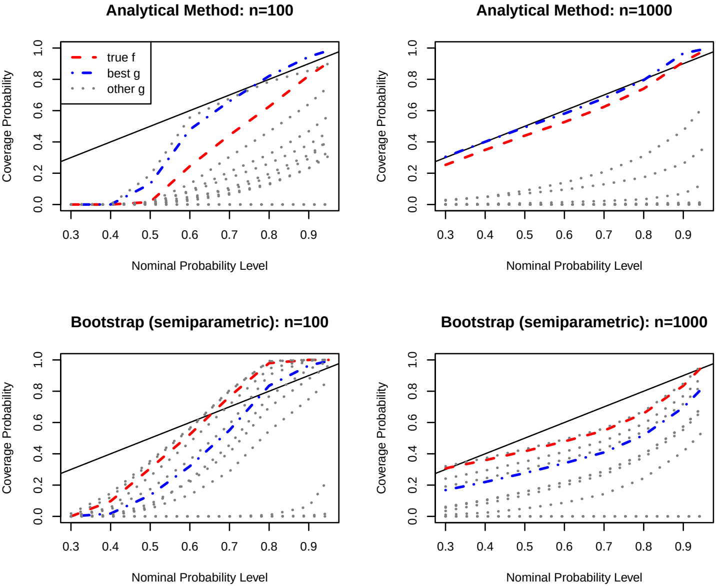
$$CE(\alpha) = F^{-1}_{X_{(1)}}(1-\alpha)$$

(This is the Analytical method)

Three choices (at least):

- **Parametric:**
  - resample from model with lowest GIC;
  - calculate quantile analytically from selected model;
  - biased if incorrect model identified...
- **Nonparametric:**
  - resample from raw data;
  - calculate quantile empirically;
  - low accuracy...
- **Semiparametric:**
  - resample from raw data to calculate Λ;
  - get quantile as in Analytical method;
  - best!

IID Case Simulations: Coverage Probabilities of CE(α) (14 models)



- Observations: $(y_1, \boldsymbol{x}_1), \ldots, (y_n, \boldsymbol{x}_n) \sim Y|\boldsymbol{x}$, in GLM setup.
- Closely parallels iid case; main modification is that:

$$y \mapsto y|\boldsymbol{x}$$

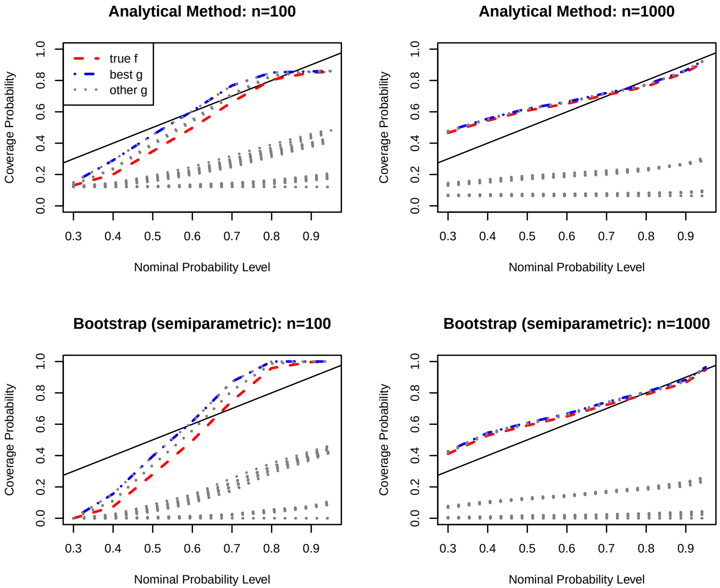
- X_j is GIC for j-th candidate model.
- Under Lindeberg-Feller regularity conditions have analogous asymptotics:

$$\sqrt{n}\begin{pmatrix} X_j - \Delta(\boldsymbol{\theta}_0^{(j)}) \\ X_k - \Delta(\boldsymbol{\theta}_0^{(k)}) \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \Lambda\right),$$

with

$$\lambda_{j,k} = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathrm{Cov}_F\left(\log g^{(j)}(Y|\boldsymbol{x}_i; \boldsymbol{\theta}_0^{(j)}), \log g^{(k)}(Y|\boldsymbol{x}_i; \boldsymbol{\theta}_0^{(k)})\right).$$

Regression Case Simulations: Coverages of CE(α) (5 predictors, all subset models)



- Observations: $y_1, \ldots, y_n \sim Y \sim f$, an ARMA$(p,q)$ model with:

$$\Sigma(\boldsymbol{\theta}) = \text{ACVF matrix}.$$

- Candidate models: $\{g_{\boldsymbol{\theta}}\}$, where $\boldsymbol{\theta} = (\phi_1, \ldots, \phi_p, \psi_1, \ldots, \psi_q, \sigma)^\intercal$.
- (Gaussian) log-likelihood for $g_{\boldsymbol{\theta}}$ is:

$$\ell_n(\boldsymbol{\theta}) = -\frac{1}{2}\log|\Gamma(\boldsymbol{\theta})| - \frac{1}{2}\boldsymbol{y}^\intercal\Gamma^{-1}(\boldsymbol{\theta})\boldsymbol{y}, \qquad \Gamma(\boldsymbol{\theta}) = \text{ACVF matrix}.$$

- KL discrepancy: between f and $g_{\boldsymbol{\theta}}$

$$\Delta_n(\widehat{\boldsymbol{\theta}}) = \frac{1}{n}\log|\Gamma(\widehat{\boldsymbol{\theta}})| + Q_n(\widehat{\boldsymbol{\theta}}), \qquad Q_n(\widehat{\boldsymbol{\theta}}) = \frac{1}{n}\boldsymbol{y}^\intercal\Gamma^{-1}(\widehat{\boldsymbol{\theta}})\boldsymbol{y}$$

- Since $\log|\Gamma(\widehat{\boldsymbol{\theta}})|/n \xrightarrow{p} 0$, just need to focus on $QF = Q_n(\widehat{\boldsymbol{\theta}})$...

Under mild regularity conditions (Mikosch, 1991) we have:

Theorem

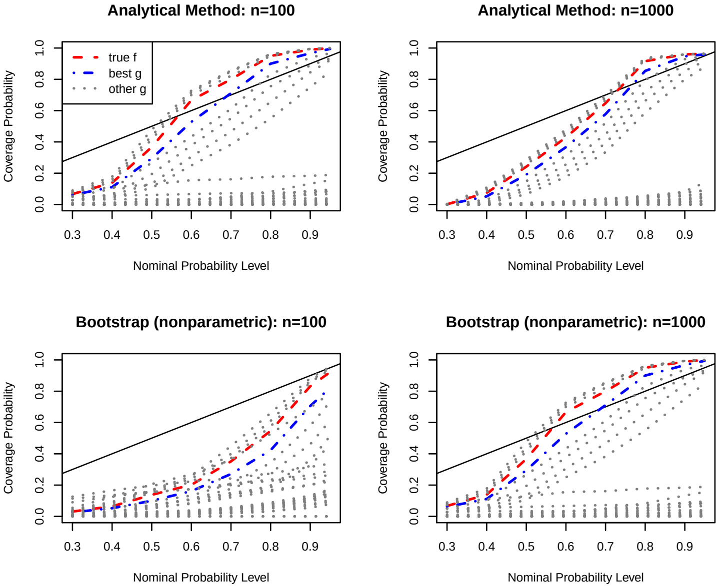
$$\sqrt{n}\begin{pmatrix} Q_n(\widehat{\boldsymbol{\theta}}^{(j)}) - \Delta(\boldsymbol{\theta}_0^{(j)}) \\ Q_n(\widehat{\boldsymbol{\theta}}^{(k)}) - \Delta(\boldsymbol{\theta}_0^{(k)}) \end{pmatrix} \xrightarrow{d} \mathcal{N}(0, \Lambda)$$

where (j, k)-th element of Λ is given by:

$$\lambda_{j,k} = \frac{2}{n} Tr\left(\Sigma(\boldsymbol{\theta}_0)\Gamma_j^{-1}(\boldsymbol{\theta}_0^{(j)})\Sigma(\boldsymbol{\theta}_0)\Gamma_k^{-1}(\boldsymbol{\theta}_0^{(k)})\right).$$

- Problem: must identify $f \leftrightarrow \Sigma(\boldsymbol{\theta})$ to estimate Λ!!!
- Solution: Identify f with lowest GIC model. . .

Time Series Case Simulations: Coverages of CE(α) ($f \sim$ AR(1), $g \sim$ ARMA(p, q))



Rank	AIC	Model	Confidence Envelope (Quantile)
1	4.9354	Gamma3	
2	4.9355	Gamma	50% (4.9365)
3	4.9374	Burr	
4	4.9418	Weibull	
5	4.9418	Weibull3	
6	4.9438	Gumbel	60% (4.9463), 70% (4.9561)
7	4.9631	LogNormal3	80% (4.9658)
8	4.9658	Lognormal	
9	4.9695	Logistic3	
10	4.9706	Logistic	
11	4.9776	Inverse Gaussian	90% (4.9854), 95% (5.0037)
12	5.0532	Kappa	
13	5.1380	Gpareto	
14	7.1821	F	

Revisit: Regression Models for Strength of Concrete

Rank	BIC	Model	Confidence Envelope (Quantile)
1	7.5703	{123458}	50% (7.5705)
2	7.5714	{12348}	
3	7.5766	{1234568}	60% (7.5723), 70% (7.5742), 80% (7.5754)
4	7.5767	{123478}	
5	7.5768	{1234578}	90% (7.5772)
6	7.5779	{123468}	95% (7.5790)
7	7.5799	{12345678}	
8	7.5827	{1234678}	
9	7.5867	{1235678}	
10	7.6142	{123678}	

Revisit: Models for Wolfer Sunspot Numbers

Rank	AICc	Model	Confidence Envelope (Quantile)
1	836.6704608	$ARMA(8,0)$	80% (836.967)
2	837.8329215	$ARMA(3,5)$	90% (837.445)
3	838.1403819	$ARMA(5,3)$	
4	838.2688009	$ARMA(5,5)$	95% (838.508)
5	838.844269	$ARMA(4,4)$	
6	838.8467446	$ARMA(8,1)$	
7	838.9254096	$ARMA(9,0)$	
8	839.0495172	$ARMA(9,1)$	
9	839.8700548	$ARMA(3,2)$	
10	840.0848802	$ARMA(9,2)$	
11	840.5555259	$ARMA(11,1)$	
12	840.9585325	$ARMA(4,2)$	
13	841.1186244	$ARMA(8,2)$	
14	841.1439068	$ARMA(4,5)$	
15	841.1898953	$ARMA(10,1)$	
16	841.1954145	$ARMA(11,0)$	
17	841.3298421	$ARMA(9,13)$	
18	841.7005855	$ARMA(12,13)$	
19	841.7189501	$ARMA(12,0)$	
20	842.4678861	$ARMA(3,3)$	

- Pivot cdf of minimum GIC to do PMSI...
- Ex: $f \sim \text{Gamma}(1, b)$ and $g \sim \text{Gamma}(2, \beta)$.
- Let: $\hat{x}$ =observed min GIC; $\hat{\beta}$ =MLE of β.
- $(1-\alpha)100\%$ CI for b: solve for b_L & b_U in:

$$F_{X_{(1)}}(\hat{x}; b_L, \hat{\beta}) = 1 - \frac{\alpha}{2}, \qquad \text{and} \qquad F_{X_{(1)}}(\hat{x}; b_U, \hat{\beta}) = \frac{\alpha}{2}.$$

- Compare with Wald CI: $\hat{b} \pm z_{1-\alpha/2}\hat{b}/\sqrt{n}$
- Simulations show (b_L, b_U) always wider!

Thank You!