Supplement for "Statistical inference for generalized additive models: simultaneous confidence corridors and variable selection"TEST

Shuzhuan Zheng, Rong Liu, Lijian Yang[1] and Wolfgang Härdle

Soochow University, University of Toledo, Soochow University, Humboldt-Universität zu Berlin

Simulation

We compare the coverage frequency of SCC with the VOT (Volume of Tube) method in the setup of the simulation 1 in Wiesenfarth et al. (2012). The model is

$$Y_i = c + \sum\nolimits_{l=1}^{3} m_l\left(X_{il}\right) + \varepsilon_i,$$

with X_{il} are independent and uniformly distributed over $[0,1]$ and $\varepsilon_i \sim N\left(0, \sigma^2\right)$. The three functions are:

$$\begin{aligned} m_1(x) &= \sin^2\left\{2\pi\left(x-0.5\right)\right\} \\ m_2(x) &= \frac{6}{10}\beta_{30,17}(x) + \frac{4}{10}\beta_{3,11}(x) \\ m_3(x) &= x\left(1-x\right) \end{aligned}$$

where $\beta_{r,s}(x) = \Gamma(r+s)\left\{\Gamma(r)\Gamma(s)\right\}^{-1} x^{r-1}(1-x)^{s-1}$. We have not run 2000 replications as in Section 5, as the results in Wiesenfarth et al. (2012) are based on 100 replications.

Table 3 shows that the performance of our proposed SCC is quite similar to the VOT method Wiesenfarth et al. (2012).

σ	n	m_1		m_2		m_3	
		SCC	VOT	SCC	VOT	SCC	VOT
	300	0.93	0.94	0.92	0.94	0.93	0.95
0.33	600	0.94	0.95	0.94	0.94	0.94	0.95
	1000	0.95	0.96	0.94	0.95	0.95	0.96
	300	0.92	0.94	0.92	0.93	0.93	0.94
0.50	600	0.93	0.94	0.94	0.95	0.93	0.95
	1000	0.94	0.95	0.94	0.95	0.95	0.96
	300	0.90	0.93	0.86	0.88	0.92	0.95
1.00	600	0.93	0.95	0.92	0.93	0.93	0.96
	1000	0.94	0.94	0.93	0.94	0.95	0.97

Table 3 The 95% SCC and VOT (Volume of Tube) coverage frequency for $m_l(x), l = 1,2,3$ from 100 replications.

Next we compare SBK and COSSO in terms of probability prediction for model (23) in Section 5 with $d = 10$. The BIC procedure is used first to

[1] Corresponding author email: yanglijian@suda.edu.cn

selection variables and then predicted probability is computed based on SBK estimation. Similar steps are taken for the COSSO method. The first 90% of data are used as training set to compute conditional probability of $Y = 1$ for the remaining 10%, and the AR as in (25) are calculated for comparison. The average and standard deviation of AR from 2000 replications are given in Table 4, which shows the SBK method has higher average AR in all cases, and smaller standard deviation except in 3 cases.

d	r	n	average AR		std. dev.	
			SBK	COSSO	SBK	COSSO
		250	0.4466	0.3836	0.1471	0.1607
	0	500	0.4415	0.3860	0.1128	0.1255
		1000	0.4415	0.3869	0.0749	0.0873
		250	0.5213	0.5198	0.1356	0.1260
10	0.5	500	0.5182	0.5115	0.0846	0.0908
		1000	0.5250	0.5115	0.0576	0.0616
		250	0.5496	0.5399	0.1110	0.1053
	0.9	500	0.5444	0.5421	0.0681	0.0618
		1000	0.5426	0.5334	0.0502	0.0517

Table 4 The average and standard deviation of AR for SBK and COSSO from 500 replications.

Proof of Theorem 2

Prior to proving Theorem 2, we restate Proposition 1 for any index set $S \subset \{1, 2, \dots d\}$.

Denote by $\boldsymbol{\lambda} = (\lambda_0, \lambda_{J,l})^T_{1\le J\le N+1, 1\le l\le d}$ an arbitrary vector. For any S, $N_S = 1 + (N+1)\,\#(S)$, denote

$$\boldsymbol{\lambda}_S = (\lambda_0, \lambda_{J,l})^T_{1\le J\le N+1, l\in S} \in \mathbb{R}^{N_S}, \mathbf{B}_S(\mathbf{x}) = \{1, B_{J,l}(x_l)\}^T_{1\le J\le N+1, l\in S},$$

and with slight abuse of notations

$$\widehat{L}_S(\boldsymbol{\lambda}_S) = \widehat{L}_S\left\{\boldsymbol{\lambda}_S^T \mathbf{B}_S(\mathbf{x})\right\} = n^{-1}\sum\nolimits_{i=1}^n \left[Y_i \boldsymbol{\lambda}_S^T \mathbf{B}_S(\mathbf{X}_i) - b\left\{\boldsymbol{\lambda}_S^T \mathbf{B}_S(\mathbf{X}_i)\right\}\right] \tag{34}$$

whose maximizer is $\widehat{m}_S = \widehat{\boldsymbol{\lambda}}_S^T \mathbf{B}_S(\mathbf{x})$.

Proposition 2 *Under Assumptions (A1)-(A5) and (A7), for m_S given in (21), $\widehat{m}_S$ in (34), as $n \to \infty$,*

$$\|m_S - \widehat{m}_S\|_{2,n} + \|m_S - \widehat{m}_S\|_2 = \mathcal{O}_{a.s.}\left(N^{1/2} n^{-1/2} \log n\right)$$

and $\|m_S - \widehat{m}_S\|_\infty = \mathcal{O}_{a.s.}\left(N n^{-1/2} \log n\right)$.

Next, we consider two cases "underfitting" and "overfitting" for the index set S to establish Theorem 2.

Definition: if $S \supset S_0$ and $S \neq S_0$, then S overfits, while S is underfitting if $S_0 \cap S \neq S_0$ with S_0 given in Theorem 2. We shall show that $\lim_{n\to\infty} P\left(\mathrm{BIC}_S - \mathrm{BIC}_{S_0} > 0\right) = 1$ in both situations.

Proof **I**: **overfitting, i.e.,** $S \supset S_0$ **and** $S \neq S_0$. Let $\boldsymbol{\lambda}_{S_0 S} = \left\{\boldsymbol{\lambda}_{S_0}, (\lambda_{J,l'})^T_{1\le J\le N+1, l'\in S\setminus S_0}\right\}$, $\widehat{\boldsymbol{\lambda}}_{S_0 S} = \left\{\widehat{\boldsymbol{\lambda}}_{S_0}, (\lambda_{J,l'})^T_{1\le J\le N+1, l'\in S\setminus S_0}\right\}$ with $\lambda_{J,l'} \equiv 0$ and $\widehat{\boldsymbol{\lambda}}_S$ (or $\widehat{\boldsymbol{\lambda}}_{S_0}$) as the MLE of (34) (or when $S = S_0$). Note that $\widehat{L}_S\left(\widehat{\boldsymbol{\lambda}}_{S_0 S}\right) = \widehat{L}_{S_0}\left(\widehat{\boldsymbol{\lambda}}_{S_0}\right)$.

Using Taylor's expansion, $\exists$ a vector $\widetilde{\boldsymbol{\lambda}}_S$ between $\widehat{\lambda}_S$ and $\widehat{\boldsymbol{\lambda}}_{S_0 S}$, i.e., $\widetilde{\boldsymbol{\lambda}}_S = \mathbf{t}\widehat{\boldsymbol{\lambda}}_S + (\mathbf{I}_{Ns} - \mathbf{t})\,\widehat{\boldsymbol{\lambda}}_{S_0 S}$ with a $N_s \times N_s$ diagonal matrix $\mathbf{t}$ whose diagonal elements are in $[0,1]$ s.t.

$$\widehat{L}_{S_0}\left(\widehat{\boldsymbol{\lambda}}_{S_0}\right) - \widehat{L}_S\left(\widehat{\boldsymbol{\lambda}}_S\right) = \widehat{L}_S\left(\widehat{\boldsymbol{\lambda}}_{S_0 S}\right) - \widehat{L}_S\left(\widehat{\boldsymbol{\lambda}}_S\right)$$
$$= \left(\widehat{\boldsymbol{\lambda}}_{S_0 S} - \widehat{\boldsymbol{\lambda}}_S\right)^T \bigtriangledown \widehat{L}_S\left(\widehat{\boldsymbol{\lambda}}_S\right) + \frac{1}{2}\left(\widehat{\boldsymbol{\lambda}}_{S_0 S} - \widehat{\boldsymbol{\lambda}}_S\right)^T \bigtriangledown^2 \widehat{L}_S\left(\widetilde{\boldsymbol{\lambda}}_S\right)\left(\widehat{\boldsymbol{\lambda}}_{S_0 S} - \widehat{\boldsymbol{\lambda}}_S\right).$$

Since $\bigtriangledown \widehat{L}_S\left(\widehat{\boldsymbol{\lambda}}_S\right) = 0$ and $\bigtriangledown^2 \widehat{L}_S\left(\widetilde{\boldsymbol{\lambda}}_S\right)$ is given in (29), for $\widetilde{m}_S = \widetilde{\boldsymbol{\lambda}}_S^T \mathbf{B}_S(\mathbf{x})$, one has

$$\widehat{L}_{S_0}\left(\widehat{\boldsymbol{\lambda}}_{S_0}\right) - \widehat{L}_S\left(\widehat{\boldsymbol{\lambda}}_S\right) =$$
$$\frac{1}{2n}\sum_{i=1}^n b''\left\{\widetilde{\boldsymbol{\lambda}}_S^T \mathbf{B}_S(\mathbf{X}_i)\right\}\left\{\left(\widehat{\boldsymbol{\lambda}}_{S_0 S} - \widehat{\boldsymbol{\lambda}}_S\right)^T \mathbf{B}_S(\mathbf{X}_i)\right\}\left\{\left(\widehat{\boldsymbol{\lambda}}_{S_0 S} - \widehat{\boldsymbol{\lambda}}_S\right)^T \mathbf{B}_S(\mathbf{X}_i)\right\}^T$$
$$= -(2n)^{-1}\sum\nolimits_{i=1}^n b''\left\{\widetilde{m}_S(\mathbf{X}_i)\right\}\left\{\widehat{m}_{S_0}(\mathbf{X}_i) - \widehat{m}_S(\mathbf{X}_i)\right\}^2. \tag{35}$$

Now since $m = m_{S_0} = m_S \in M_{S_0} \subset M_S$, Proposition 2 implies that

$$\left\|\widehat{m}_{S_0} - m\right\|_{2,n} = \mathcal{O}_{a.s.}\left(N^{1/2} n^{-1/2} \log n\right),$$
$$\left\|\widehat{m}_S - m\right\|_{2,n} = \mathcal{O}_{a.s.}\left(N^{1/2} n^{-1/2} \log n\right).$$

Thus

$$\left\|\widehat{m}_{S_0} - \widehat{m}_S\right\|^2_{2,n} = \mathcal{O}_{a.s.}\left(N n^{-1} \log^2 n\right). \tag{36}$$

Similarly, one has $\left\|\widetilde{m}_S - m\right\|_\infty = o_{a.s.}(1)$, which warrants for large n that $\widetilde{m}_S \in \Theta$ with Θ given in Assumption (A.2), so (35) implies that

$$0 \geq \widehat{L}_{S_0}\left(\widehat{\boldsymbol{\lambda}}_{S_0}\right) - \widehat{L}_S\left(\widehat{\boldsymbol{\lambda}}_S\right) \geq -\frac{C_b}{2}\left\|\widehat{m}_{S_0} - \widehat{m}_S\right\|^2_{2,n}. \tag{37}$$

As a result, BIC given in (22) shows that

$$\mathrm{BIC}_S - \mathrm{BIC}_{S_0} = 2\left\{\widehat{L}_{S_0}\left(\widehat{\boldsymbol{\lambda}}_{S_0}\right) - \widehat{L}_S\left(\widehat{\boldsymbol{\lambda}}_S\right)\right\} + \frac{N_S - N_{S_0}}{n}\log^3 n$$
$$\geq -C_b\left\|\widehat{m}_{S_0} - \widehat{m}_S\right\|^2_{2,n} + (N+1)\, n^{-1}\log^3 n,$$

which implies by (36) that $\lim_{n\to\infty} P\left(\mathrm{BIC}_S - \mathrm{BIC}_{S_0} > 0\right) = 1$.

II: underfitting, i.e., $S_0 \cap S \neq S_0$.

Let $S' = S_0 \cup S$ and denote by $\widehat{\boldsymbol{\lambda}}_{S_0}$, $\widehat{\boldsymbol{\lambda}}_S$ and $\widehat{\boldsymbol{\lambda}}_{S'}$ the MLEs in (34) for S_0, S and S', respectively. Since S' overfits S_0, similarly to (37), one has

$$0 \geq \widehat{L}_{S_0}\left(\widehat{\boldsymbol{\lambda}}_{S_0}\right) - \widehat{L}_{S'}\left(\widehat{\boldsymbol{\lambda}}_{S'}\right) \geq -\frac{C_b}{2}\left\|\widehat{m}_{S_0} - \widehat{m}_{S'}\right\|_{2,n}^2, \tag{38}$$

Define a set $\Theta_{S'} = \left\{\boldsymbol{\lambda}_{S'} : \boldsymbol{\lambda}_{S'}^T \mathbf{B}_{S'}\left(\mathbf{X}_i\right) \in \Theta, 1 \leq i \leq n\right\}$. which is compact and convex in $\mathbb{R}^{N_S}$. By definition,

$$\max_{1\leq i\leq n}\left|\widehat{\boldsymbol{\lambda}}_{S'}^T \mathbf{B}_{S'}\left(\mathbf{X}_i\right) - m\left(\mathbf{X}_i\right)\right| \leq \left\|\widehat{m}_{S'} - m\right\|_\infty = \mathcal{O}_{a.s.}\left(Nn^{-1/2}\log n\right),$$

so for large n, with probability approaching 1, $\widehat{\boldsymbol{\lambda}}_{S'} \in \Theta_{S'}$, so Proposition 1 ensures that, with probability approaching 1, the Hessian matrix $\nabla^2 \widehat{L}_{S'}\left(\widehat{\boldsymbol{\lambda}}_{S'}\right) \leq -c_b c_V \mathbf{I}_{N_{S'}}$, while $\nabla \widehat{L}_{S'}\left(\widehat{\boldsymbol{\lambda}}_{S'}\right) = 0$ and $\nabla^2 \widehat{L}_{S'}\left(\boldsymbol{\lambda}_{S'}\right) \leq \mathbf{0}$, $\forall\ \boldsymbol{\lambda}_{S'}$. Thus, with probability approaching 1, there exists a constant $c_1 > 0$ such that

$$\widehat{L}_{S'}\left(\boldsymbol{\lambda}_{S'}\right) - \widehat{L}_{S'}\left(\widehat{\boldsymbol{\lambda}}_{S'}\right) \leq \begin{cases} -2^{-1} c_b c_V \left\|\boldsymbol{\lambda}_{S'} - \widehat{\boldsymbol{\lambda}}_{S'}\right\|^2, & \text{if } \boldsymbol{\lambda}_{S'} \in \Theta_{S'} \\ \max\limits_{\boldsymbol{\lambda}_{S'} \in \partial\Theta_{S'}} \widehat{L}_S\left(\boldsymbol{\lambda}_S\right) - \widehat{L}_{S'}\left(\widehat{\boldsymbol{\lambda}}_{S'}\right) \leq -c_1, & \text{otherwise} \end{cases}. \tag{39}$$

Next, define a new vector $\widehat{\boldsymbol{\lambda}}_{SS'} = \left\{\widehat{\boldsymbol{\lambda}}_S, \left(\lambda_{J,l'}\right)_{1\leq J\leq N+1, l' \in S'\setminus S}^T\right\}$ with $\lambda_{J,l'} \equiv 0$ and note that $\widehat{\boldsymbol{\lambda}}_{SS'}^T\ \mathbf{B}_{S'}\left(\mathbf{x}\right) \equiv \widehat{m}_S\left(\mathbf{x}\right), \widehat{\boldsymbol{\lambda}}_{S'}^T\ \mathbf{B}_{S'}\left(\mathbf{x}\right) \equiv \widehat{m}_{S'}\left(\mathbf{x}\right)$, so applying Lemma A.5 of Wang and Yang (2007) (the proof of which, although reviewed and accepted as an integral part of the article, was published in a separate online version), there exists a constant $C_0 > 0$ such that

$$\left\|\widehat{\boldsymbol{\lambda}}_{SS'}^T - \widehat{\boldsymbol{\lambda}}_{S'}^T\right\|^2 \geq C_0^{-1}\left\|\left(\widehat{\boldsymbol{\lambda}}_{SS'}^T - \widehat{\boldsymbol{\lambda}}_{S'}^T\right)\mathbf{B}_{S'}\left(\mathbf{x}\right)\right\|_2^2,$$

hence

$$\left\|\widehat{\boldsymbol{\lambda}}_{SS'} - \widehat{\boldsymbol{\lambda}}_{S'}\right\|^2 \geq C_0^{-1}\left\|\widehat{m}_S - \widehat{m}_{S'}\right\|_2^2.$$

Applying Proposition 2 entails that

$$\left|\left\|\widehat{m}_S - \widehat{m}_{S'}\right\|_2^2 - \left\|m_S - m_{S'}\right\|_2^2\right| = \mathcal{O}_{a.s.}\left(Nn^{-1}\log^2 n\right)$$

while the definitions of underfitting and overfitting lead to

$$\left\|m_S - m_{S'}\right\|_2^2 = \left\|m_S - m\right\|_2^2 = c_S > 0$$

and thus

$$\left\|\widehat{\boldsymbol{\lambda}}_{SS'} - \widehat{\boldsymbol{\lambda}}_{S'}\right\|^2 \geq C_0^{-1} c_S + \mathcal{O}_{a.s.}\left(Nn^{-1}\log^2 n\right). \tag{40}$$

Note next that

$$\widehat{L}_S\left(\widehat{\boldsymbol{\lambda}}_S\right) - \widehat{L}_{S'}\left(\widehat{\boldsymbol{\lambda}}_{S'}\right) = \widehat{L}_{S'}\left(\widehat{\boldsymbol{\lambda}}_{SS'}\right) - \widehat{L}_{S'}\left(\widehat{\boldsymbol{\lambda}}_{S'}\right) \tag{41}$$

which according to (39), is bounded by

$$\leq \begin{cases} -2^{-1}c_b c_V \left\|\widehat{\boldsymbol{\lambda}}_{SS'} - \widehat{\boldsymbol{\lambda}}_{S'}\right\|^2, & \text{if } \widehat{\boldsymbol{\lambda}}_{SS'} \in \Theta_{S'} \\ \max\limits_{\boldsymbol{\lambda}_{S'} \in \partial\Theta_{S'}} \widehat{L}_S\left(\boldsymbol{\lambda}_S\right) - \widehat{L}_{S'}\left(\widehat{\boldsymbol{\lambda}}_{S'}\right) \leq -c_1, & \text{otherwise} \end{cases}$$

which is, according to (40), bounded by

$$\begin{aligned} &\leq \max\left(-2^{-1}c_b c_V C_0^{-1} c_S, -c_1\right) + \mathcal{O}_{a.s.}\left(Nn^{-1}\log^2 n\right) \\ &= -c_2 + \mathcal{O}_{a.s.}\left(Nn^{-1}\log^2 n\right), \text{ for a constant } c_2 > 0. \end{aligned}$$

The above bound, together with (36), (38) and (41) lead to $\widehat{L}_{S_0}\left(\widehat{\boldsymbol{\lambda}}_{S_0}\right) - \widehat{L}_S\left(\widehat{\boldsymbol{\lambda}}_S\right)$

$$\begin{aligned} &= \left\{\widehat{L}_{S_0}\left(\widehat{\boldsymbol{\lambda}}_{S_0}\right) - \widehat{L}_{S'}\left(\widehat{\boldsymbol{\lambda}}_{S'}\right)\right\} - \left\{\widehat{L}_S\left(\widehat{\boldsymbol{\lambda}}_S\right) - \widehat{L}_{S'}\left(\widehat{\boldsymbol{\lambda}}_{S'}\right)\right\} \\ &\geq c_2 + \mathcal{O}_{a.s.}\left(Nn^{-1}\log^2 n\right). \end{aligned} \tag{42}$$

Finally, (42) implies that

$$\mathrm{BIC}_S - \mathrm{BIC}_{S_0} = 2\left\{\widehat{L}_{S_0}\left(\widehat{\boldsymbol{\lambda}}_{S_0}\right) - \widehat{L}_S\left(\widehat{\boldsymbol{\lambda}}_S\right)\right\} + \frac{N_S - N_{S_0}}{n}\log^3 n \geq c_2 + o_p(1),$$

and thus $\lim_{n\to\infty} P\left(\mathrm{BIC}_S - \mathrm{BIC}_{S_0} > 0\right) = 1$.