

SZIMMETRIA ÉS KONFIDENCIA

CSÁJI BALÁZS CSANÁD

A cikkben áttekintjük az SPS (Sign-Perturbed Sums) becslési módszert, amely a zaj szimmetriáját kihasználva egzakt, nem aszimptotikus konfidenciatartományokat tud konstruálni regressziós modellek pontbecslései köré.

1. Bevezetés

Regressziós modellek becslése zajos megfigyelési adatokból (pl. idősorokból) fontos statisztikai probléma, amely számos területen előkerül, például a rendszer identifikációban, jelfeldolgozásban, gépi tanulásban és a pénzügyi matematikában.

A standard módszerek – például az előrejelzési hiba-, maximum likelihood- és korrelációs módszer [5] – tipikusan *pontbecsléseket* szolgáltatnak. Gyakorlati szempontból is fontos kérdés, hogy mennyire bízhatunk a kapott becslésben, amelyre egy lehetséges válasz, hogy megadott valószínűségű *konfidenciatartományt* építünk a becslés köré. Az ilyen halmazok konstrukciójának klasszikus módja az, ha a pontbecslés határeloszlását hívjuk segítségül [5]. Az aszimptotikus eredményeken alapuló megközelítések azonban nem nyújtanak szigorú garanciákat véges mintaszám esetén, hacsak nem teszünk erős statisztikai feltevéseket a rendszerről.

A dolgozat célja a nemrég kidolgozott SPS (Sign-Perturbed Sums) módszer [3] bemutatása, amely minimális statisztikai feltevésekkel képes regressziós modellek pontbecslései köré egzakt, nem aszimptotikus konfidenciahalmazokat építeni.

Az SPS egy *félparametrikus* becslési módszer, mivel egy parametrikus (akár dinamikus) modelltől indul ki, de a folyamatot hajtó zajra nézve nem tételez fel paraméterezést. Először az egyszerű lineáris regresszió esetére mutatjuk be a módszert, majd kiterjesztjük dinamikus rendszerekre is a konstrukciót.

2. Konfidenciahalmazok lineáris regressziós problémákhoz

Adott egy bemenet-kiment párokból álló minta, $\mathcal{D}_n \triangleq \{(\varphi_1, Y_1), \dots, (\varphi_n, Y_n)\}$,

$$Y_t \triangleq \varphi_t^T \theta^* + N_t,$$

$t \in \{1, \dots, n\}$, ahol Y_t a kimenet (magyarázott változó), φ_t a bemenet (magyarázó változó, regresszor) és N_t pedig a (nem megfigyelhető) zaj, a t -edik megfigyelés esetén. A cél az ismeretlen, „igazi” θ^* paraméter becslése. Feltesszük, hogy a bemenetek, $\{\varphi_t\} \subset \mathbb{R}^d$, determinisztikusak, $\theta^* \in \mathbb{R}^d$ konstans, és a zaj, $\{N_t\}$, független, szimmetrikus (a nulla körül) eloszlású¹ (valós értékű) valószínűségi változókból áll; azaz N_t és $-N_t$ eloszlása megegyezik minden t -re. Az egyszerűség kedvéért szintén feltesszük, hogy $n > d$ és a $\Phi \triangleq [\varphi_1, \dots, \varphi_n]$ mátrix teljes (sor) rangú.

2.1. Legkisebb négyzetek becslés- és konfidenciaellipszoidjai

Az egyik standard pontbecslés a jól ismert *legkisebb négyzetek* (LS) becslés,

$$\hat{\theta}_n \triangleq \arg \min_{\theta \in \mathbb{R}^d} \mathcal{V}(\theta | \mathcal{D}_n) = \arg \min_{\theta \in \mathbb{R}^d} \frac{1}{2} \|Y - \Phi^T \theta\|_2^2,$$

ahol $Y \triangleq [Y_1, \dots, Y_n]^T$. A *normálegyenlet* megoldásával megkapható $\hat{\theta}_n$, azaz

$$\nabla_{\theta} \mathcal{V}(\hat{\theta}_n | \mathcal{D}_n) = \Phi \Phi^T \hat{\theta}_n - \Phi Y = 0,$$

aminek a fenti feltételek mellett az analitikus megoldása $\hat{\theta}_n = (\Phi \Phi^T)^{-1}(\Phi Y)$.

Fontos kérdés, hogy mennyire megbízható a kapott pontbecslés. Erre egy válasz, ha tudunk például egy megadott $p \in (0, 1)$ valószínűségű $\hat{\Theta}_{\mathcal{D}_n, p}$ konfidencia-tartományt készíteni $\hat{\theta}_n$ köré, azaz amelyre $\hat{\theta}_n \in \hat{\Theta}_{\mathcal{D}_n, p}$, és $\mathbb{P}(\theta^* \in \hat{\Theta}_{\mathcal{D}_n, p}) \geq p$.

Ilyen halmazok konstrukciójára egy standard módszer [5], ha kihasználjuk, hogy az LS-becslés (átskálázott) hibája aszimptotikusan Gauss-eloszlású, azaz

$$\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{d} \mathcal{N}(0, \sigma^2 R^{-1}), \quad \text{ahogy } n \rightarrow \infty,$$

amely fennáll pl., ha korlátosak a regresszorok, és létezik egy olyan pozitív definit R mátrix, amely az $R_n \triangleq \frac{1}{n} \Phi_n \Phi_n^T$ határértéke², valamint $\{N_t\}$ független, azonos eloszlású (f.a.e.) változókból áll, amelyekre $\mathbb{E}[N_t] = 0$, és $\mathbb{E}[N_t^2] = \sigma^2$, $0 < \sigma^2 < \infty$.

A pontbecslés határeloszlásának felhasználásával egy adott p valószínűséghez $\hat{\theta}_n$ középpontú (közelítő) konfidenciaellipszoid konstruálható,

$$\tilde{\Theta}_{n, p} \triangleq \left\{ \theta \in \mathbb{R}^d : (\theta - \hat{\theta}_n)^T R_n (\theta - \hat{\theta}_n) \leq \frac{q \hat{\sigma}_n^2}{n} \right\}, \quad (1)$$

¹A függetlenség gyengíthető, a szimmetria a kritikus az SPS-hez. Számos nevezetes eloszlás lehet ilyen, például Gauss, Laplace, Cauchy, Bernoulli, Student t, egyenletes.

²Itt kivételesen – a határérték miatt – expliciten kiírtuk, hogy Φ függ n -től.

ahol $p = F_{\chi^2(d)}(q)$, itt $F_{\chi^2(d)}$ a d szabadságfokú χ^2 -eloszlás eloszlásfüggvénye, és $\hat{\sigma}_n^2$ a zaj varianciájának becslése az LS-megoldás reziduálisainak segítségével,

$$\hat{\sigma}_n^2 \triangleq \frac{1}{n-d} \sum_{t=1}^n (Y_t - \varphi_t^T \hat{\theta}_n)^2.$$

Ekkor nyilván $\hat{\theta}_n \in \tilde{\Theta}_{n,p}$, valamint $\mathbb{P}(\theta^* \in \tilde{\Theta}_{n,p}) \approx p$. Az így kapott konfidencia halmazok azonban véges minták esetén nem garantáltak, és csak heurisztikus megközelítésnek tekinthetők; kis mintaszám esetén tipikusan pontatlanok [5].

2.2. SPS lineáris regresszió esetén

Most rátérünk az SPS-módszer [3] ismertetésére, amellyel minimális statisztikai feltevésekkel véges minták esetén is garantált konfidenciatartományokat konstruálhatunk az LS-becslés köré. Első közelítésben tekinthetünk az SPS-re úgy, mint egy hipotézisvizsgálatra, amely egy adott θ paraméter esetén azt a nullhipotézist vizsgálja, hogy $\theta = \theta^*$, a $\theta \neq \theta^*$ alternatív hipotézissel szemben.

A *bootstrap*- és *Monte-Carlo*-tesztek alap gondolatához hasonlóan abból indulunk majd ki, hogy ha $\theta = \theta^*$, akkor egyrészt (i) megkaphatjuk a zajváltozók pontos értékét a rendszer „invertálásával”, valamint (ii) a zaj eloszlásának bizonyos regularitását kihasználva (a jelen esetben a szimmetriát) alternatív mintákat generálhatunk, amelyek statisztikailag „hasonlóan viselkednek” majd, mint az eredeti minta. Ha azonban $\theta \neq \theta^*$, akkor a rendszer invertálásával kapott zajbecslések torzítottak lesznek, és így az alternatív minták statisztikailag „eltérően viselkednek” majd, mint az eredeti minta. Azt, hogy mennyire „hasonlóan viselkednek” a generált (perturbált) értékek az eredeti mintához viszonyítva, a minták skalár értékű kiértékelése után egy rangteszttel döntjük majd el.

Első lépésként bevezetünk m (véletlenített) kvadratikusan $\mathbb{R}^d \rightarrow \mathbb{R}$ függvényt,

$$Z_i(\theta) \triangleq \|R_n^{-\frac{1}{2}} \Phi \Lambda_i (Y - \Phi^T \theta)\|_2^2,$$

$i \in \{0, 1, \dots, m-1\}$, ahol $\Lambda_0 \triangleq I$, és $\Lambda_i \triangleq \text{diag}(\alpha_{i,1}, \dots, \alpha_{i,n})$, ha $i \neq 0$; $\{\alpha_{i,t}\}$ független, azonos eloszlású Rademacher-valószínűségű változók³; a $\text{diag}(\cdot)$ függvény pedig egy diagonális mátrixot képez az argumentumaiból. A Z_0 -t *referenciafüggvénynek*, míg a többi $\{Z_i\}$ függvényt *előjelperturbált* függvényeknek nevezzük.

Vegyük észre a kapcsolatot a $\{Z_i\}$ függvények definíciója és a költségfüggvény gradiense, $\nabla_{\theta} \mathcal{V}$, között. A $\{Z_i\}$ függvényeket interpretálhatjuk úgy, hogy \mathcal{V} gradiensében a reziduálisok előjelét véletlenítjük Rademacher-változók segítségével, majd a kapott vektor „nagyságát” kiértékeljük egy súlyozott norma segítségével.

³Azaz 1 és -1 értékeket vehetnek fel, mindegyiket $1/2$ valószínűséggel.

Ha $\theta = \theta^*$, akkor $Y - \Phi^T \theta^* = N$, ahol $N = [N_1, \dots, N_n]^T$, és a szimmetriafeltevésből tudjuk, hogy N és $\Lambda_i N$ eloszlása minden i -re megegyezik. Ekkor

$$Z_0(\theta^*) = \|R_n^{-\frac{1}{2}} \Phi N\|_2^2 \stackrel{d}{=} \|R_n^{-\frac{1}{2}} \Phi \Lambda_i N\|_2^2 = Z_i(\theta^*),$$

de a $\{Z_i(\theta^*)\}$ valószínűségi változók természetesen nem teljesen függetlenek. Be lehet látni azonban, hogy *feltételesen* f.a.e. tulajdonságúak az $\{|N_t|\}$ által generált σ -algebrára nézve. Ekkor viszont *felcserélhetőek* is, és így minden lehetséges rendezésük⁴, $Z_{i_0}(\theta^*) \prec \dots \prec Z_{i_{m-1}}(\theta^*)$, ugyanolyan $1/m!$ valószínűséggel áll elő.

Ha azonban $\theta \neq \theta^*$, akkor már ez a felcserélhetőségi tulajdonság nem áll fenn, és $Z_0(\theta)$ egyre nagyobb valószínűséggel fogja dominálni a többi $\{Z_i(\theta)\}_{i \neq 0}$ változót, ahogy egyre távolabb kerülünk az igazi paramétertől, azaz ahogy $\|\theta^* - \theta\|_2 \rightarrow \infty$.

Ahhoz, hogy össze tudjuk mérni a referenciafüggvény értékét az előjelperturbált függvényekével, szükségünk lesz a referenciafüggvény *normalizált rangjára*,

$$\mathcal{R}(\theta) \triangleq \frac{1}{m} \left(1 + \sum_{i=1}^{m-1} \mathbb{I}(Z_0(\theta) \prec Z_i(\theta)) \right),$$

ahol $\mathbb{I}(\cdot)$ egy indikátorfüggvény; tehát 1, ha az argumentuma (egy formula) igaz, és 0 máskülönben. Tegyük fel, hogy $p = 1 - q/m$ alakban írható, ahol $0 < q < m$, és q, m egész számok. Ekkor az SPS-teszt elfogadja a nullhipotézist, ha $\mathcal{R}(\theta) \leq p$, és elutasítja, ha $\mathcal{R}(\theta) > p$. Mivel mind m , a referencia- és előjelperturbált függvények száma, mind q szabad paraméterek (a felhasználó által meghatározott), így tetszőleges racionális⁵ $p \in (0, 1)$ valószínűséghez készíthető egy SPS-teszt.

A fentiek szellemében az SPS-konfidenciatartományt így definiálhatjuk:

$$\widehat{\Theta}_{n,p} \triangleq \{ \theta \in \mathbb{R}^d : \mathcal{R}(\theta) \leq p \}.$$

Bebizonyítható [3], hogy az így kapott (véletlenített) konfidenciahalmaz *egzakta*,

$$\mathbb{P}(\theta^* \in \widehat{\Theta}_{n,p}) = p.$$

A konfidenciahalmaz egzaktága annak ellenére fennáll, hogy nem használtuk ki a zaj konkrét eloszlását, sőt, még azt is megengedtük, hogy minden megfigyelésre különböző eloszlású zaj hasson, amelyeknek akár végtelen varianciája is lehet.

Az SPS konfidenciahalmazai *csillagkonvexek*, ahol az LS-beclés egy csillagközpont [3]. Tehát minden $\theta \in \widehat{\Theta}_{n,p}$ és $\beta \in [0, 1]$ -re, $\beta \theta + (1 - \beta) \hat{\theta}_n \in \widehat{\Theta}_{n,p}$.

⁴A „ \prec ” egy szigorú teljes rendezés, amit a standard „ $<$ ”-ből úgy kapunk, hogy egyenlő értékek esetén véletlenszerűen döntjük el, hogy melyiket tekintjük kisebbnek; formális def. lásd [3].

⁵További véletlenítéssel könnyen kiterjeszthető a teszt irracionális valószínűségekre is, azonban ennek elhanyagolható a gyakorlati jelentősége, ezért ismertetésétől eltekintünk.

Az SPS-konfidenciahalmazok *erősen konzisztensek* [2], azaz aszimptotikusan egy valószínűséggel semmilyen hamis paraméterértéket nem tartalmaznak; a halmaz θ^* körüli tetszőleges kicsi gömb belsejébe kerül, ahogy a mintaszám növekszik,

$$\mathbb{P} \left(\bigcup_{k=1}^{\infty} \bigcap_{n=k}^{\infty} \{ \hat{\Theta}_n \subseteq B_\varepsilon(\theta^*) \} \right) = 1,$$

minden $\varepsilon > 0$, ahol $B_\varepsilon(\theta^*) \triangleq \{ \theta \in \mathbb{R}^d : \|\theta - \theta^*\|_2 \leq \varepsilon \}$. A konzisztencia gyenge plusz feltevések mellett fennáll, pl. mind a zaj szórásnégyzete, mind a regresszorok normája tarthat a végtelenhez, ha a növekedésük bizonyos ráta alatt marad [2].

Hatékony *külső approximációs ellipszoid* is konstruálható az SPS-halmazokhoz:

$$\hat{\Theta}_{n,p} \subseteq \{ \theta \in \mathbb{R}^d : (\theta - \hat{\theta}_n)^T R_n (\theta - \hat{\theta}_n) \leq r^* \},$$

amelyek gyorsan (polinomiális időben) – *szemidefinit programozási* feladatok megoldásával – számolhatóak. Vegyük észre, hogy az így kapott konfidenciaellipszoidoknak ugyanaz lesz a középpontja (ti. az LS-beclés), és az alakját meghatározó mátrixa, mint (1) esetén, csak az ellipszoidok sugara fog különbözni. Azonban, míg a határeloszláson alapuló ellipszoidok csak heurisztikák, az SPS-halmaz külső approximációján alapuló ellipszoidok garantált konfidenciával rendelkeznek [3].

3. Konfidenciahalmazok dinamikus rendszerekhez

Az alapvető célkitűzés az SPS-módszer megalkotásakor az volt, hogy *dinamikus rendszerek* pontbecslései köré tudjunk nem aszimptotikus garanciákkal rendelkező konfidenciahalmazokat készíteni. A lineáris regressziós eset egy természetes általánosítása, ha *általános lineáris (dinamikus) rendszereket* [5] vizsgálunk, azaz

$$Y_t \triangleq G(z^{-1}; \theta^*) U_t + H(z^{-1}; \theta^*) N_t,$$

ahol G és H (stabil) lineáris szűrők, H stabilan invertálható, z^{-1} a késeltetés („lag”) operátor, $\theta^* \in \mathbb{R}^d$ az „igazi” paraméter, $G(0; \theta^*) = 0$, $H(0; \theta^*) = 1$, $\{U_t\}$ a (megfigyelt) bemenetek, $\{N_t\}$ a (nem megfigyelt) független és szimmetrikus zaj, a bemenetek és a zaj függetlenek, valamint a rendszer kezdeti állapota ismert [4].

Ekkor az *előrejelzési hibák* vektora⁶ $\hat{\varepsilon}(\theta) \triangleq [\hat{\varepsilon}_1(\theta), \dots, \hat{\varepsilon}_n(\theta)]^T$, ahol $\hat{\varepsilon}_t(\theta) \triangleq H^{-1}(z^{-1}; \theta) (Y_t - G(z^{-1}; \theta) U_t)$ minden t -re; valamint teljesül, hogy $\hat{\varepsilon}_t(\theta^*) = N_t$.

Egy tipikus költségfüggvény az előrejelzési hibák négyzetösszege [5], azaz

$$\mathcal{V}(\theta | \mathcal{D}_n) = \frac{1}{2} \hat{\varepsilon}^T(\theta) \hat{\varepsilon}(\theta), \quad \text{ekkor} \quad \Psi(\hat{\theta}_n) \hat{\varepsilon}(\hat{\theta}_n) = 0,$$

⁶Tegyük fel, hogy pont annyi adatunk van, hogy n hibatagot kiszámolhassunk.

ahol a $\hat{\theta}_n$ az *előrejelzési hibabecslés*, $\Psi(\theta) \triangleq [\psi_1(\theta), \dots, \psi_n(\theta)]$, és $\psi_t(\theta)$ pedig a t -edik előrejelzési hiba gradiense. Ismert, hogy ez a gradiens $\psi_t(\theta) = W(z^{-1}; \theta) Y_t + K(z^{-1}; \theta) U_t$ alakban is írható, ahol W és K (vektorértékű) lineáris szűrők [5].

Ha megpróbáljuk ugyanazt a konstrukciót használni, mint a . fejezetben, az sajnos erre az általános esetre nem fog működni, a $Z_0(\theta^*)$ -nak *nem ugyanaz lesz az eloszlása*, mint a többi $\{Z_i(\theta^*)\}_{i \neq 0}$ -nek, így az igazi paraméterhez tartozó függvényértékek felcserélhetőségéhez vezető érvelés sem marad érvényben [4].

A probléma a rendszer dinamikusságából fakad; azaz abból, hogy ha a zajt perturbáljuk, az – a lineáris regressziós esettől eltérően – hatással van az előrejelzési hibák gradiensére is. A megoldást alternatív kimeneti trajektóriák generálása jelenti, amelyeket a perturbált előrejelzési hibákkal hajtunk meg:

$$\bar{Y}_{i,t}(\theta) \triangleq G(z^{-1}; \theta) U_t + H(z^{-1}; \theta) (\alpha_{i,t} \hat{\varepsilon}_t(\theta)),$$

$i \in \{1, \dots, m-1\}$, és a jelölés egyszerűsítése miatt bevezetjük, hogy minden t -re $\bar{Y}_{0,t}(\theta) \triangleq Y_t$. Az alternatív trajektóriákkal perturbált gradienseket számolhatunk,

$$\bar{\psi}_{i,t}(\theta) \triangleq W(z^{-1}; \theta) \bar{Y}_{i,t}(\theta) + K(z^{-1}; \theta) U_t,$$

$i \in \{0, \dots, m-1\}$, amelyekkel eljutunk az SPS működőképes általánosításához:

$$Z_i(\theta) \triangleq \|Q_i^{-\frac{1}{2}}(\theta) \Psi_i(\theta) \Lambda_i \hat{\varepsilon}(\theta)\|_2^2,$$

ahol $\Psi_i(\theta) \triangleq [\bar{\psi}_{i,1}(\theta), \dots, \bar{\psi}_{i,n}(\theta)]$, és $Q_i(\theta) \triangleq \frac{1}{n} \Psi_i(\theta) \Psi_i^T(\theta)$. Bebizonyítható, hogy az ezekkel a függvényekkel konstruált SPS-halmazok már *egzakt* konfidenciával rendelkeznek [4]. A konstrukció kiterjeszthető *visszacsatolt* (pl. szabályozóval rendelkező) rendszerekre [4], valamint *nemlineáris* (pl. GARCH) modellekre is [1].

4. Konklúzió és nyitott problémák

A dolgozatban röviden bemutatottuk az SPS-módszert, amely minimális statisztikai feltevésekkel képes regressziós modellek adott pontbecslései köré véges minták esetén is garantált – tipikusan egzakt – konfidenciatartományokat készíteni.

Ugyan az SPS-módszer viselkedését már elég jól ismerjük lineáris regressziós problémákon, de még sok nyitott kérdés van vele kapcsolatban például dinamikus rendszerek esetén. Ilyen kérdés az SPS-teszt másodfajú hibájának jellemzése és hatékony külső approximációk konstruálása különböző dinamikus modellekhez.

Köszönetnyilvánítás

Csáji B. Cs. munkáját az MTA Bolyai János Kutatási Ösztöndíja, valamint az ED_18-2-2018-0006. és 125698. számú NKFIH-pályázatok támogatásával végezte.

Hivatkozások

- [1] CSÁJI, B. CS.: *Score Permutation Based Finite Sample Inference for Generalized Autoregressive Conditional Heteroskedasticity (GARCH) Models*, in: Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS), Cadiz, Spain, pp. 296-304 (2016).
- [2] CSÁJI, B. CS., CAMPI, M. C., AND WEYER, E.: *Strong Consistency of the Sign-Perturbed Sums Method*, in: Proceedings of the 53rd IEEE Conference on Decision and Control, Los Angeles, California, pp. 3352-3357 (2014). DOI: [10.3182/20140824-6-ZA-1003.02704](https://doi.org/10.3182/20140824-6-ZA-1003.02704)
- [3] CSÁJI, B. CS., CAMPI, M. C., AND WEYER, E.: *Sign-Perturbed Sums: A New System Identification Approach for Constructing Exact Non-Asymptotic Confidence Regions in Linear Regression Models*, IEEE Trans. on Signal Proc., Vol. **63** No. **1**, pp. 169-181 (2015). DOI: [10.1109/TSP.2014.2369000](https://doi.org/10.1109/TSP.2014.2369000)
- [4] CSÁJI, B. CS. AND WEYER, E.: *Closed-Loop Applicability of the Sign-Perturbed Sums Method*, in: Proceedings of the 54th IEEE Conference on Decision and Control, Osaka, Japan, pp. 1441-1446 (2015). DOI: [10.1109/CDC.2015.7402413](https://doi.org/10.1109/CDC.2015.7402413)
- [5] LJUNG, L.: *System Identification: Theory for the User*, 2nd edn., Prentice-Hall, Upper Saddle River (1999). DOI: [10.1109/MRA.2012.2192817](https://doi.org/10.1109/MRA.2012.2192817)



Csáji Balázs Csanád a SZTAKI tudományos főmunkatársa, 1976-ban született. Első diplomáját programtervező matematikusként szerezte az ELTE-TTK-n 2001-ben, majd filozófia szakos bölcsész diplomát szerzett az ELTE-BTK-n 2006-ban. Tanulmányai alatt 3-5 hónapos részképzésekben vett részt az Eindhoveni Műszaki Egyetemen (Hollandia, 2001), a British Telecomnál (Nagy Britannia, 2002) és a Johannes Kepler Egyetemen (Ausztria, 2003). PhD-fokozatát az ELTE Informatikai Karán védte meg 2008-ban. Doktorálása után a Louvaini Katolikus Egyetemen (Belgium) volt posztdoktori kutató,

majd 2009-től a Melbourne-i Egyetemen (Ausztrália) dolgozott, ahonnan 2013-ban tért haza. Csáji Balázs Csanád eredményeit több díjjal jutalmazták, például elnyerte az Ausztrál Kutatási Tanács (ARC) Discovery Early Career Researcher Award (DECRA) díját, valamint az MTA Matematikai Tudományok Osztályának Gyires Béla-díját. Összesen 70 referált tudományos cikk szerzője, kutatási területe a gépi tanulásban és rendszer identifikációban fellépő sztochasztikus modellek valószínűségelméleti és statisztikai vizsgálata.

CSÁJI BALÁZS CSANÁD

Magyar Tudományos Akadémia
Számítástechnikai és Automatizálási Kutatóintézet
1111 Budapest, Kende utca 13-17.
csaji.balazs@sztaki.mta.hu

SYMMETRY AND CONFIDENCE

BALÁZS CSANÁD CSÁJI

In this article we briefly overviewed the SPS (Sign-Perturbed Sums) method which was recently developed by Marco Campi, Erik Weyer and Balázs Csanád Csáji. The main draw card of SPS is that it can construct non-asymptotically guaranteed confidence regions for parameters of regression models under minimal statistical assumptions. The fundamental assumption of the SPS method is that the noises affecting the system are distributed symmetrically about zero, but their distributions may change over time.

We started with investigating classical linear regression problems. We recalled the standard least-squares (LS) estimate with its confidence ellipsoids, whose construction is based on the asymptotic Gaussianity of the estimate. We argued that such ellipsoids do not come with rigorous finite sample guarantees and are imprecise for small samples.

Then, we presented the construction of SPS confidence regions for linear regression problems. We highlighted that these regions have (i) exact confidence probabilities; are (ii) star convex with the LS estimate as a star center; are (iii) strongly consistent, that is, for every ball around the true parameter, the confidence regions are asymptotically (as the sample size tends to infinity) almost surely remain inside the ball (thus every false parameter value will be eventually excluded); and (iv) efficient ellipsoidal outer-approximations can be constructed for them by solving semidefinite programming problems. The SPS ellipsoids have the same center (i.e., the LS estimate) and kernel matrix as the classical ellipsoids of the LS theory, only their radii are different. However, while the ellipsoids of the classical theory (based on limiting distributions) are just heuristics for finite samples, the SPS ellipsoids have rigorous non-asymptotic guarantees.

Finally, we generalized the SPS construction for general linear (dynamical) systems. We discussed that a straightforward generalization would not work, as the noises affecting the system and the outputs are not independent for dynamical systems. It was shown that the proper generalization should be built using alternative output trajectories, which are constructed using perturbed prediction error sequences. The resulting SPS confidence regions also have exact confidence probabilities, and the construction can be generalized for closed-loop and certain non-linear systems (such as GARCH models), as well.

There are several open problems concerning the SPS method, especially for its construction for dynamical systems: for example, building efficient outer-approximations for various types of dynamical systems and analyzing the type II errors of SPS tests.

Keywords: confidence regions, resampling methods, linear regression, time series.

Mathematics Subject Classification (2000): 62F25, 66G09, 62J05, 62M10.